Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise Visual dos Dados Educacionais Voltada para o Estudo de Gênero nos Cursos de Computação da Universidade de Brasília

Luiza A. Hansen, Lucas M. Chagas

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr. Vinicius Borges

Coorientadora
Prof.a Dr.a Maristela Terto de Holanda

Brasília
2018

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise Visual dos Dados Educacionais Voltada para o Estudo de Gênero nos Cursos de Computação da Universidade de Brasília

Luiza A. Hansen, Lucas M. Chagas

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr. Vinicius Borges (Orientador)
CIC/UnB

Prof.a Dr.a Aletéia Patrícia Favacho de Araújo
CIC/UnB

Prof. Dr. Thiago Paulo Faleiros
CIC/UnB

Prof. Dr. Edison Ishikawa
Coordenador do Bacharelado em Ciência da Computação

Brasília, 24 de dezembro de 2018

Dedicatória

Luiza: Dedico este trabalho a minha família e amigos que apoiaram nas horas mais difíceis e sorriram comigo nas horas mais alegres. Sem eles, este trabalho e muitos dos meus sonhos não se realizariam.
Lucas: Dedico este trabalho aos meus amigos, que me ajudaram em todos os trabalhos, estudos, e todas outras atividades acadêmicas. Dedico também aos meus pais, que sempre mostraram apoio nas minhas decisões e estiveram do meu lado em todos os momentos. Dedico a minha irmã, que sempre foi uma excelente companheira, minha inspiração e uma grande guerreira. Por fim, dedico este trabalho ao meu primo, que foi quem fez eu escolher este curso, que me ajudou e foi meu exemplo, e agora está lá em cima, me vigiando e protegendo.

Agradecimentos

Gostaríamos de agradecer aos orientadores, Professora Doutora Maristela Holanda e Professor Doutor Vinícius Borges, que auxiliaram na construção da nossa pesquisa e se disponibilizaram para ajudar sempre que necessário. Ao SIGRA pelo material disponibilizado, essencial para a realização do estudo

Luiza: Agradeço, com carinho especial, a todas as pessoas que me apoiaram e auxiliaram a revisar o texto: Tereza Cristina de Melo Aguiar, Inácio Cauduro Hansen, Wânia Mara de Melo Aguiar, Rafael Aires de Alencar Lucas da Silva e Rômulo Feijão Filho

Lucas: Gostaria de agradecer ao departamento de CiC e aos professores que tive durante a graduação, todos sendo de fundamental importância para a minha formação acadêmica. Por fim, gostaria de agradecer a minha dupla, Luiza Aguiar Hansen, que sem ela, esse e tantos outros trabalhos não seriam possíveis.

Resumo

O número de mulheres em cursos de tecnologia vem diminuindo com o passar dos anos, chegando em 2016 a menos de 20% do total do corpo estudantil do Departamento de Ciência da Computação da Universidade de Brasília. A utilização de visualizações auxilia na tomada de decisão para promover a entrada e a permanência de alunas nos cursos e, consequentemente, aumentar o número de mulheres no mercado de trabalho em áreas de tecnologia.

Este trabalho emprega técnicas de visualização para analisar e identificar padrões no perfil de meninas nos cursos de graduação da área de tecnologia. Por isso,foram utilizadas técnicas de redução de dimensionalidade (PCA e t-SNE), Mapas de Calor e Gráficos de Coordenadas Paralelas para o processo de análise visual de dados, considerando a situação das estudantes em relação à UnB (ativas, desligadas ou graduadas). As visualizações obtidas revelaram que os dados possuem natureza não linear, sendo possível agrupar as meninas de acordo com a forma de saída. Neste trabalho, foi evidenciada a correlação existente entre as variáveis, sendo analisada mais profundamente a associação entre os períodos de entrada e de saída na Universidade, e a forma de saída desta.

Palavras-chave: Visualização de Informações, Dados Educacionais, Mulheres na Tecnologia

Abstract

The number of women in technology courses has been decreasing over the years, reaching less than 20% of the total student body of the Department of Computer Science of the University of Brasília in the year of 2016. The use of visualization helps in the decision making, to promote the entrance and the permanency of female students in courses, thus increasing the number of women in the labor market in technology areas.

This work uses visualization techniques to analyze and identify profile patterns in girls of undergraduate courses in the technology field. Dimensionality reduction techniques (PCA and t-SNE), HeatMap and Parallel Coordinates Graphics were used for the visual data analysis process, considering the students situation in relation to UnB (active, disconnected or graduated). The visualizations obtained revealed that the data have a non-linear nature, making it possible to join girls in groups according to their form of leaving the university. In this work, the existing correlations between variables were evidenced, being deeper analyzed the association between periods of entrance and leaving in the university and the way of leaving.

Keywords: Information Visualization, Educational Data, Women in Technology

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

A visualização é uma ação da visão humana que permite o reconhecimento de conjuntos de informações. Até o século XVII, a visualização de dados consistiu em diagramas geométricos, mapas e tabelas de posição das estrelas. Com o passar do tempo surgiram novas técnicas, também utilizadas para mapear clima, relevo, probabilidade, economia, entre outros, além de inovar em cores e novos padrões [2]. O advento do computador e das tecnologias altamente interativas impulsionou o desenvolvimento de novas técnicas gráficas e novos métodos de visualização multidimensional, apoiados na facilidade de armazenamento e manipulação de dados.

Segundo Freitas et al. [3], "*as técnicas de visualização de informações procuram representar graficamente dados de um determinado domínio de aplicação, de modo que a representação visual gerada explore a capacidade de percepção do homem e este, a partir das relações espaciais exibidas, interprete e compreenda as informações apresentadas e, finalmente, gere novos conhecimentos*". De modo geral, essas técnicas auxiliam a análise e a compreensão de um conjunto de dados, por meio da geração de representações gráficas e dos mecanismos de interação, que evidenciam a observação de características ou padrões no conjunto. Assim, as técnicas de visualização de informações podem ser usadas para [4]:

- Entender informações rapidamente, sendo possível enxergar grandes quantidades de dados de modo compreensível e coeso;
- Identificar últimas tendências, facilitando a descoberta de valores atípicos;
- Identificar relações e padrões, que devido a apresentação gráfica, indicam coerência em grandes volumes de dados;

- Envolver o usuário e transmitir a mensagem rapidamente, por meio do impacto causado pelos diagramas, gráficos e outras representações visuais geradas.

A mineração de dados, em conjunto com as técnicas de visualização, vem sendo empregada na Educação, para o desenvolvimento de métodos na exploração de conjuntos de dados coletados em ambientes pedagógicos [5]. Desta forma, é possível entender o desempenho dos alunos e moldar o ambiente para fornecer melhores condições de aprendizado, além de viabilizar a identificação de formas mais eficazes de ensino e de abordagem, que proporcionam melhores benefícios educacionais.

Uma dessas abordagens educacionais seria avaliar o comportamento das mulheres em cursos na área de computação. Estima-se que o Brasil possua, em média, apenas 18% de concluintes do sexo feminino em tais cursos, de acordo com estatísticas construídas a partir dos dados do INEP, do censo da educação superior de 2016, realizado pela Sociedade Brasileira de Computação (SBC) [6]. Especificamente na Universidade de Brasília, foi observado que os cursos do Departamento da Ciência da Computação apresentam uma porcentagem muito baixa de participação das meninas em relação aos meninos, chegando a menos de 20% no ano de 2016 [7].

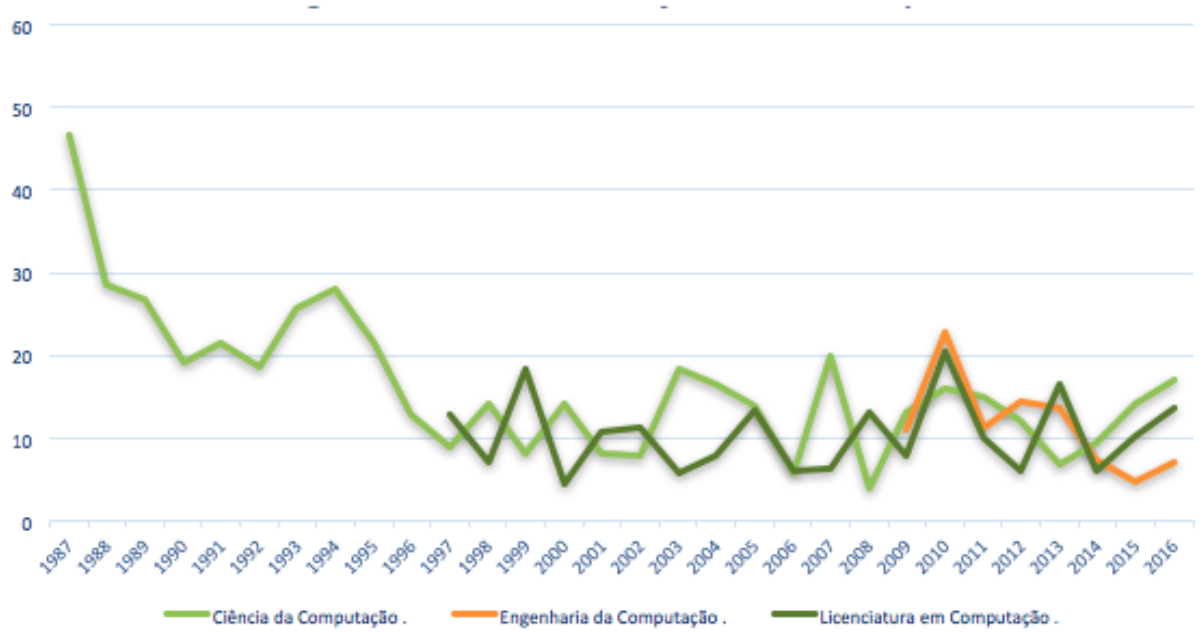


Figura 1.1: Relação dos alunos do sexo feminino por ano na UnB

Comparada a outros cursos, a Ciência da Computação mostrou ser, no passado, uma área bastante promissora para mulheres, chegando a ter mais participação feminina do que em cursos como medicina e física [8]. Entretanto, a partir da década de 80, observa-se que a presença de alunas nos cursos de tecnologia vem sofrendo grande queda, ao passo

que outros cursos têm registrado um aumento nestes números [1], conforme ilustrado na Figura 1.2.

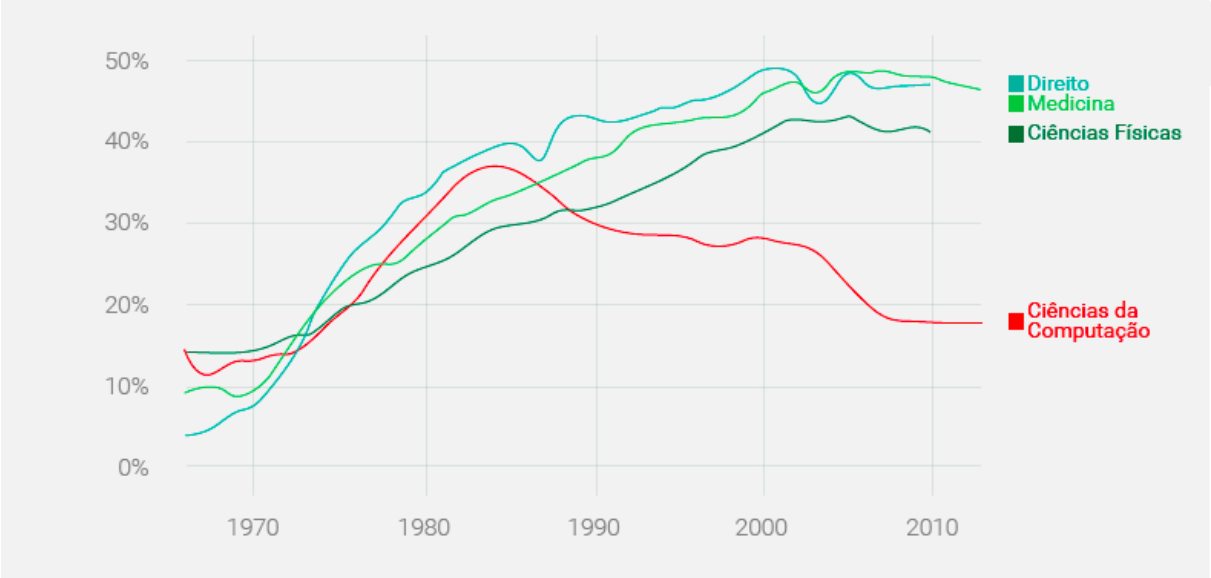


Figura 1.2: Porcentagem de especialistas nos cursos de Direito, Medicina, Ciências Físicas e Ciência da Computação ao longo dos anos. Adaptação de [1].

A palavra Computação, do latim *computatĭo*, é um substantivo feminino que tem como definições o ato ou o efeito de computar, o conjunto de conhecimentos e técnicas referentes ao uso de computadores e o processamento automático de dados [9]. Na história da computação, a influência de grandes mulheres foi de extrema importância para a evolução da área.

Ada Lovelace, matemática e escritora inglesa, ficou conhecida no século XIX por seu trabalho, em conjunto com o cientista Charles Babbage, na construção da Máquina Analítica, que resultou no primeiro algoritmo criado na história, muito antes da existência de máquinas que pudessem processá-lo [10][11]. A Irmã Mary Kenneth Keller, pioneira no estudo de Ciência da Computação, foi a primeira mulher a concluir um doutorado na área, em 1965, na Universidade do Wisconsin-Madison. Além de defender as mulheres dentro dos cursos de computação, Irmã Mary Kenneth Keller também lutou pela inclusão de tecnologia nas escolas [12]. Uma equipe formada por seis mulheres desenvolveu, em 1946, o primeiro computador totalmente eletrônico e programável, denominado ENIAC [13]. Durante a produção desse supercomputador, elas também foram responsáveis pela criação de diversos protocolos utilizados até hoje, pelo teclado numérico e pelo primeiro sistema informatizado para o censo americano, além de influenciarem sistemas de "salvamento" de configurações e preferências [14].

Para alterar esse cenário de queda no Brasil, o Congresso da Sociedade Brasileira de Computação (CSBC) realiza, desde 2007, o Workshop "*Women in Information and Tech-*

nology" (WIT). A SBC tem ainda o projeto Meninas Digitais com o objetivo de discutir o tema da baixa inserção das mulheres na área de Computação [15], tendo desenvolvido diversos projetos parceiros em todas as regiões do Brasil. Além disso, órgãos do governo do Brasil, como o Ministério de Ciência, Tecnologia e Inovação (MCTI), lançaram editais para projetos de pesquisa específicos para a formação de meninas em cursos de Exatas e Computação [16]. A iniciativa privada, como o Instituto Unibanco, lançou em 2015 o Edital "Gestão Escolar para Equidade: Elas nas Exatas", que visa a reduzir o impacto da desigualdade de gênero nas Ciências Exatas [17]. Em 2018, o CNPq lançou o edital Chamada CNPq/MCTIC Nº 31/2018 - Meninas nas Ciências Exatas, Engenharias e Computação, no qual o projeto Meninas.comp da UnB foi um dos projetos vencedores deste edital.

1.2 Objetivo

O presente estudo tem como objetivo analisar o conjunto de dados das mulheres nos cursos de computação da UnB, visando encontrar padrões implícitos, como característica pessoais, acadêmicas e sociais. Para isso, serão utilizadas técnicas de visualização de informação que facilitam o entendimento e tornam as informações mais acessíveis a uma tomada de decisão.

Para alcançar o objetivo geral deste trabalho, os seguintes objetivos específicos foram evidenciados:

- Desenvolvimento do Banco de Dados utilizando informações sobre os alunos dos cursos de tecnologia da UnB;
- Elaboração de um método para análise visual de dados educacionais;
- Aplicação das técnicas definidas para extração de informações relevantes ao objetivo.

1.3 Estrutura do Documento

Este documento é composto pelos seguintes capítulos:

- Capítulo 2: Introduz os conceitos Estudo de Dados e Visualização da Informação, apresenta os fundamentos de probabilidade e estatística, bem como elucida as técnicas de visualização tradicionais, coordenadas paralelas e *heatmap*; e técnicas de visualização para dados multivariados, *PCA* e *t-SNE*;

- Capítulo 3: Apresenta artigos que analisaram a temática de mulheres na computação e a mineração de dados educacionais, assim como artigos sobre a utilização de visualização na área de educação, que auxiliaram na construção desta pesquisa;
- Capitulo 4: Justifica os critérios adotados para a escolha do conjunto de dados utilizado na pesquisa, além de descrever a limpeza desses dados, a realização da mineração de dados e o levantamento das hipóteses da pesquisa;
- Capítulo 5: Apresenta os resultados obtidos, utilizando as técnicas de visualização tradicionais e as técnicas de visualização para dados multivariados, que fundamentaram e responderam as hipóteses levantadas.
- Capítulo 6: Expõe as conclusões obtidas a partir do desenvolvimento deste trabalho e apresenta sugestões para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo revisa os fundamentos utilizados para realização das análises de perfil das meninas estudantes das áreas de computação. Inicialmente, são tratados na Seção 2.1 os princípios de banco de dados, o conceito de pirâmide de dados, a transformação dos dados em ciência e a sua execução dentro de técnicas de visualização, utilizando o modelo tabular. A Seção 2.2 discorre sobre os fundamentos de probabilidade e estatística. A Seção 2.3 conceitua visualização de informação e visualização exploratória em conjunto com seus processos, sendo apresentadas as técnicas de visualização tradicionais, e duas técnicas de visualização de dados multivariados empregadas neste trabalho.

2.1 Estudo de Dados

2.1.1 Dados

Banco de dados é uma coleção de dados inter-relacionados, representando informações de um domínio específico [18]. A sua utilização varia desde aplicações voltadas para o mercado, até aplicações acadêmicas, contemplando as áreas de telecomunicações, trânsito aéreo, finanças, dentre outras.

Entender o significado de instância e dado é necessário para compreender sua relevância no estudo de banco de dados. Instância é um conjunto de dados armazenados em um determinado instante de tempo [19]. Dados representam propriedades de objetos e eventos, podendo ser quantitativos e qualitativos [20].

O dado é mensurável, coletado, reportado e analisado, sendo possível visualizá-lo em gráficos, imagens ou qualquer ferramenta de análise. Dados brutos não carregam nenhuma significância além da sua existência, podendo ou não serem utilizáveis. Por outro lado, a informação consiste em um agrupamento de dados de forma organizada, gerando

conhecimento e inferida a partir das perguntas: Quem? O que? Quando? Quantos? [21]. A principal diferença entre dado e informação é a funcionalidade, não a estrutura [22].

2.1.2 Modelo Tabular

Um conjunto de dados $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}\}$ é definido por N instâncias multidimensionais, ou seja, cada instância $\mathbf{x_i} = \{x_{i,1}, ...x_{i,m}\}$ é descrita por m atributos, sendo interpretada como um vetor m-dimensional, logo o conjunto de dados $\mathbf{X}$ define um espaço m-dimensional.

Conjuntos de dados podem ser representados por um modelo tabular, em que as instâncias estão associadas com as linhas e os atributos com as colunas. A Tabela 2.1 ilustra esse modelo, em que, uma instância está associada a um registro e pode ser descrita com atributos caracterizados como UF de Nascimento, sexo, curso, opção, entre outros. Desta forma, os valores de uma mesma coluna possuem o mesmo tipo e os elementos da linha podem ser comparados entre si.

Tabela 2.1: Representação do modelo tabular.

UF Nascimento	Sexo	Curso	Opção
DF	Masculino	Ciência da Computação	Ciência da Computação
RN	Masculino	Ciência da Computação	Ciência da Computação
DF	Masculino	Engenharia Mecatrônica	Engenharia Mecatrônica
DF	Masculino	Ciência da Computação	Ciência da Computação
DF	Feminino	Engenharia de Redes	Engenharia de Redes
RJ	Masculino	Engenharia da Computação	Engenharia da Computação

2.1.3 Similaridade de Dados

Dentro da mineração de dados, diversas atividades utilizam princípios de dissimilaridade, ou similaridade, para realização de suas tarefas. Entre essas atividades estão: classificação; detecção de anomalia; e agrupamentos. Em muitas dessas tarefas, é necessário comparar instâncias de dados, isto é, calcular um valor que expressa o grau de dissimilaridade, ou similaridade, entre elas. O termo dissimilaridade é definido informalmente como uma medida numérica que reflete o grau de distinção entre dois dados [23]. Valores próximos de zero representam um conjunto de dados mais homogêneos e mais próximo no espaço de características. Já valores próximos a 1 caracterizam dados distintos e distantes entre si [24].

Existem diversas formas de calcular a dissimilaridade entre instâncias de dados, sendo o uso de métricas, ou funções de distância, uma das abordagens mais comuns. A distância de Minkowski é uma função generalizada que permite obter a diferença máxima entre dois

pontos, tendo $p = \infty$ [23]. A distância de Minkowski é definida pela Equação 2.1, onde p é um parâmetro que define a métrica:

$$d_M(x, y) = \left(\sum_{i=1}^{m} |x_i - y_i|^p \right)^{1/p}. \tag{2.1}$$

Algumas funções clássicas para calcular distância entre pontos em um espaço m-dimensional podem ser obtidas a partir da Distância de Minkowski. A Distância Euclidiana, também conhecida como métrica L_2, pertence à família de métricas Minkowski, onde $p = 2$. Esta métrica define um posicionamento espacial geométrico dos pontos equidistantes em relação a um ponto de referência. A Equação 2.2 descreve a distância Euclidiana:

$$d_E(x, y) = \sqrt{\left(\sum_{i=1}^{m} |x_i - y_i|^2 \right)}. \tag{2.2}$$

A Distância City-Block (Manhattan ou métrica L_1), apresentada na Equação 2.3, é obtida ao considerar $p = 1$:

$$d_{CB}(x, y) = \left(\sum_{i=1}^{m} |x_i - y_i| \right). \tag{2.3}$$

2.2 Fundamentos de Probabilidade e Estatística

Probabilidade é definida como a medida em que um determinado evento pode ocorrer. Dentro de qualquer experimento aleatório, é atribuído um valor entre 0 e 1 para identificar essa probabilidade, de forma que, quanto mais próximo de 1 - ou 100% - maiores são as chances do evento acontecer [25].

Um conceito importante dentro de probabilidade e estatística é a expectativa matemática de uma variável aleatória, também conhecida como valor esperado ou apenas expectativa. Esta medida fornece um valor único que atua como representante ou média dos valores de X, sendo chamado, por essa razão, de medida de tendência central [25].

A expectativa de X também é chamada de média de X e é determinada por μ_x, ou apenas μ. Para uma variável aleatória discreta X com valores $x_1, ..., x_n$, o valor esperado de X é definido pela Equação 2.4:

$$E(X) = x_1 P(X = x_1) + ... + x_n P(X = x_n) = \sum_{i=1}^{n} x_i P(X = x_i). \tag{2.4}$$

A expectativa matemática de variáveis aleatórias é utilizada para calcular a Covariância e a Correlação.

2.2.1 Covariância

A covariância de duas variáveis aleatórias é determinada como a medida da sua variabilidade em conjunto do seu grau de associação [26]. Covariância pode ser definida a partir da Equação 2.5:

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)], \tag{2.5}$$

em que X e Y representam as variáveis aleatórias distribuídas em conjunto e μ_x e μ_y descrevem os valores esperados.

De acordo com a Equação 2.5, pode-se deduzir que quando o valor de X estiver maior do que a média do conjunto X e o valor de Y também tender a ser maior que a média do conjunto Y, a covariância será positiva, já que as variáveis aleatórias são associadas positivamente. Em contrapartida, quando X é maior que média, mas Y tende a ser menor que a média, a covariância será negativa, sendo as variáveis aleatórias associadas negativamente.

A utilização de covariância tem aplicabilidade tanto em incorporação de dados, como em redução de dimensionalidade e extração de recursos, sendo empregado em técnicas como *Principal Component Analysis* [26].

2.2.2 Correlação

A correlação é uma técnica estatística que busca indicar a força e a direção na relação entre duas ou mais variáveis [27]. Essas variáveis estão relacionadas quando a ocorrência de mudanças no valor de uma provoca alterações no valor da outra.

O coeficiente de correlação utilizado nesta pesquisa para otimização da análise de dados será o coeficiente de correlação de Pearson, também chamado de produto-momento. Esse coeficiente de correlação é obtido pela divisão da covariância de duas variáveis pelo produto dos respectivos desvios padrão [28]. O coeficiente é avaliado em escala sem unidade e pode conter valores desde -1, passando por 0 e até +1.

A partir da Equação 2.6, é possível calcular o coeficiente de correlação de Pearson, considerando $(x, y) = (x_i, y_i)$, com $i = 1, ..., n$ para uma amostra de dados bivariados.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}}, \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}, \tag{2.6}$$

Onde r é o valor do coeficiente de correlação. Para valores de $r > 0$, tem-se que sua correlação é positiva. Para valores de $r < 0$, tem-se que sua correlação é negativa. Caso o valor de $r = 0$, então a sua correlação é neutra.

A caracterização da correlação entre positiva, negativa ou neutra define seu coeficiente de correlação. Se o coeficiente for positivo, à medida em que um valor x cresce ou decresce, o valor y varia no mesmo sentido. Se o coeficiente for negativo, à medida em que um valor x cresce, o valor y decresce e vice-versa. Por fim, se o coeficiente for neutro, não existe nenhuma associação linear [29].

Ao definir o coeficiente de correlação, é possível utilizá-lo em diversas técnicas de visualização tradicionais, gerando uma análise voltada para a relação de variação entre os fatores.

2.3 Visualização da Informação

O processo de Visualização da Informação consiste em transformar os dados apresentados em imagens reais, que possam ser visualizadas e interpretadas por seres humanos [30]. Tendo como objetivo incluir informações para o usuário por meio de imagens, figuras, estruturas gráficas e qualquer outro tipo de recurso gráfico [31], esse processo busca chamar a atenção do espectador para as características dos dados [32]. No processo de visualização é necessário saber o papel de cada técnica para cada aplicação, visto que o uso inadequado de técnicas de visualização pode gerar resultados insuficientes ou até incorretos [33].

Keim [34] apresenta que os objetivos da visualização da informação podem ser divididos em três atividades: análise exploratória, análise confirmativa e apresentação. O processo de visualização exploratória pode ser encarado como um processo de geração de hipóteses no qual as visualizações dos dados permitem ao usuário obter informações sobre os dados e criar novas hipóteses. A verificação das hipóteses também pode ser feita por meio da exploração visual de dados, por técnicas automáticas de estatística ou ainda aprendizado de máquina. Além do envolvimento direto do usuário, as principais vantagens da exploração visual de dados em técnicas de mineração automáticas, a partir de estatísticas ou aprendizado de máquina, são relacionadas à possibilidade de lidar com grande número de dados não homogêneos e ruidosos, além de ser intuitivo e não requerer conhecimento de matemática complexa, algoritmos estatísticos ou parâmetros.

2.3.1 Técnicas de Visualização Tradicionais

Coordenadas Paralelas

Coordenadas paralelas consiste em ferramenta de análise de dados multivariado, onde para qualquer N inteiro positivo, um sistema de coordenadas para o espaço m-Dimensional Euclidiano R^m é construído. No gráfico de coordenadas paralelas, cada linha é uma variável e as instâncias de dados são poli-linhas que interceptam esses eixos em uma

posição determinada pelo valor do atributo associado ao eixo [35]. Este tipo de gráfico é indicado para identificar valores discrepantes ou padrões com base em fatores de métricas relacionadas e localizar pontos de cruzamento.

A Figura 2.1 apresenta um exemplo simples de coordenadas paralelas com quatro variáveis. Pela imagem as instâncias que possuem valor mais baixo na variável "petal_length" possuem valor baixo na variável "petal_width" e valores ligeiramente mais altos na variável "sepal_width".

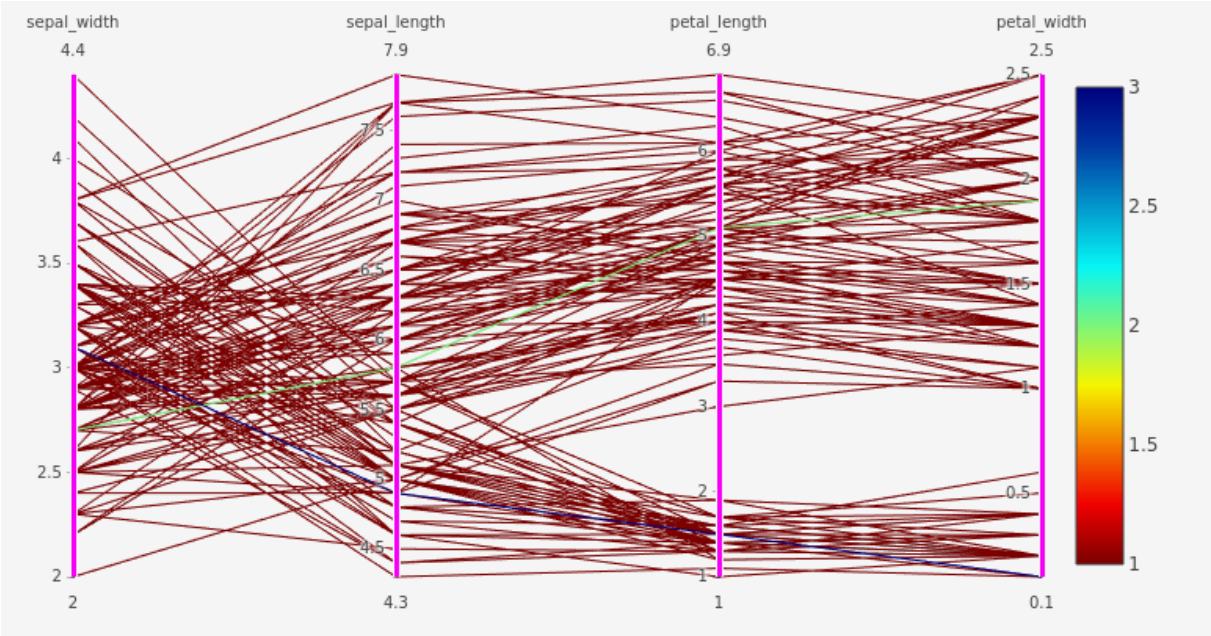


Figura 2.1: Exemplo de aplicação utilizando coordenadas paralelas.

Heatmap

Heatmap, também conhecido como mapa de calor, é uma representação da densidade geográfica de pontos em um mapa [36], onde cada cor da célula corresponde ao valor de correlação de duas determinadas variáveis [37]. Sua utilização em pesquisa e observação de padrões é datada de 1873, onde Toussaint Loua fez um levantamento sobre a distribuição da sociedade parisiense em diversos aspectos, entre eles: origem, idade, classe social, e profissão [38]. Desde então, *Heatmap* vem sendo utilizado em diversas áreas de pesquisa.

Representar os dados em formato de matriz facilita a extração do valor exato, mesmo em grandes conjuntos de dados. Da mesma forma, as cores auxiliam a rápida leitura de padrões, em escala quantitativa [39]. A utilização do *Heatmap* com coeficiente de correlação tem como intuito mostrar a intensidade da correlação entre duas variáveis.

A Figura 2.2 apresenta um exemplo de aplicação de mapa de calor. As variáveis "sepal_width", "sepal_length", "petal_length", "petal_width" estão representadas na

diagonal da tabela. As cores azuis representam uma correlação positiva, enquanto as cores vermelhas representam uma correlação negativa. Quanto mais escuro o vermelho, mais próximo do valor -1, e quanto mais escuro o azul, mais próximo do valor 1.

Por meio da figura 2.2 é possível identificar que o comprimento da pétala e o comprimento da sépala tem correlação positiva forte, já que a coluna "sepal_length" intercepta com um círculo azul escuro a linha "petal_length". A mesma análise pode ser feita com as variáveis "sepal_width" e "petal_length", que possuem correlação negativa, já que a interseção das variáveis possui um círculo vermelho.

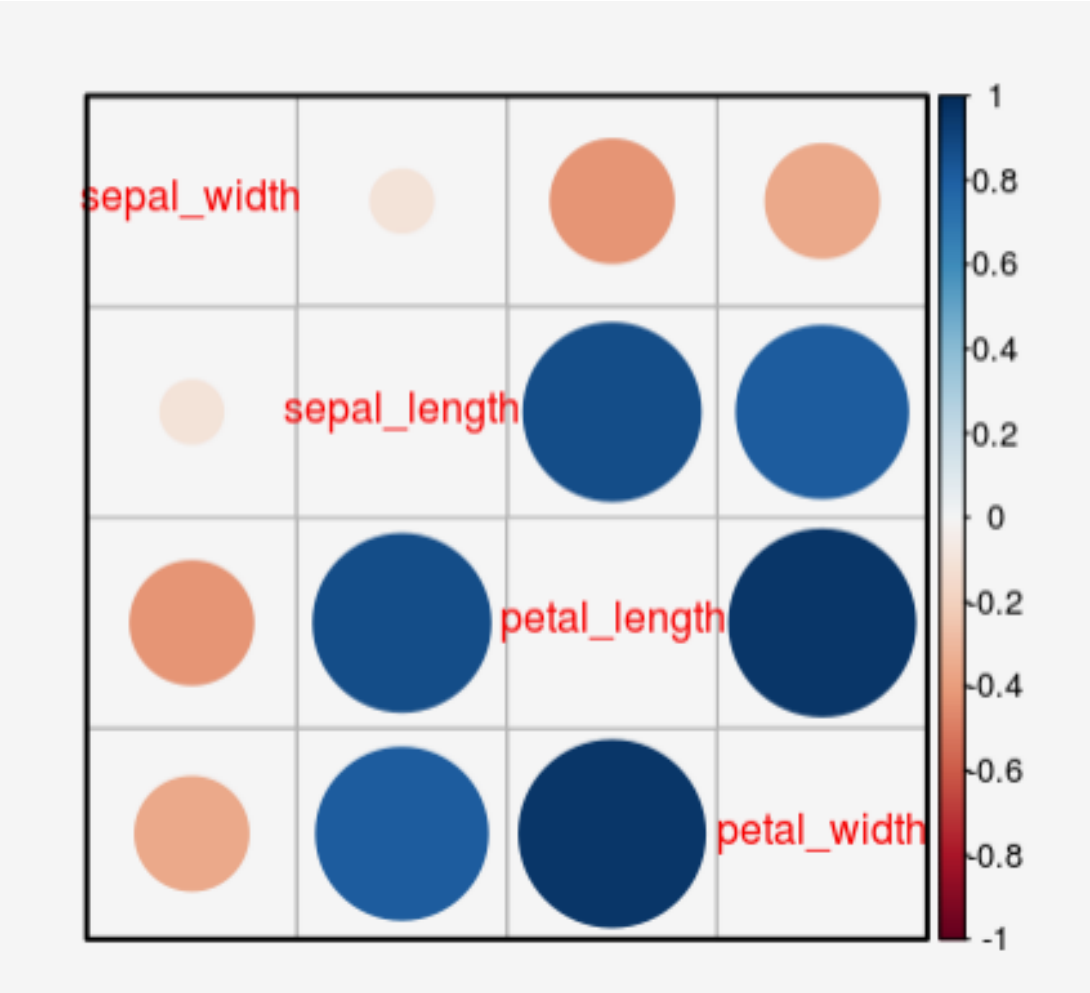


Figura 2.2: Exemplo de aplicação utilizando mapa de calor.

2.3.2 Técnicas de Visualização baseadas em Redução de Dimensionalidade

A dimensionalidade dos dados está relacionada ao número de características de padrões ou à dimensão do espaço de características. Algoritmos de aprendizado de máquina e de

processos de mineração de dados podem ter seus desempenhos afetados à medida em que a dimensionalidade dos dados aumenta [40]. Tal fato motivou a proposição de técnicas de redução de dimensionalidade, que tem como objetivo obter uma representação de baixa dimensionalidade dos dados multidimensionais, reduzindo a perda de informações, enquanto mantém as características mais relevantes dos dados. Em geral, a seleção de características reduz o custo de medição de dados, enquanto mantém a interpretação física original e as propriedades que possuíam quando foram criadas [41].

Principal Component Analysis

PCA é um método que permite a redução da dimensionalidade por meio da representação do conjunto de dados em um novo sistema de eixos, denominados componentes principais, permitindo a visualização da natureza multivariada dos dados em poucas dimensões [42].

O PCA realiza o mapeamento dos dados de espaço de alta dimensionalidade para um espaço de dimensão reduzida, no qual a variância dos dados é maximizada. Primeiramente, os dados são centralizados na origem do sistema de coordenadas e, então, calcula-se a matriz de correlação. Em seguida, realiza-se uma decomposição espectral na matriz de correlação, resultando na obtenção de autovalores e autovetores. Para alcançar o espaço reduzido com p dimensões, selecionam-se os autovetores associados aos maiores p autovalores. Para visualizar os dados do PCA em um gráfico de dispersão bidimensional, considera-se $p = 2$ [43].

PCA foi primeiramente formulado por Pearson [44], que analisava conjunto de dados como linhas e planos mais próximos dos pontos no espaço. Fisher e Mackenzie [45] o consideraram mais adequado do que a análise de variância para a modelagem de dados de resposta. Por fim, Hotelling [46] definiu o PCA utilizado atualmente. Desde então, sua utilização tem sido alvo de diversos campos de pesquisa [47].

A Figura 2.3 é uma visualização em PCA de um conjunto de dados Íris [48], contendo características de três espécies de flores: *setosa*, *virginica* e *versicolor*. Essas flores possuem características de tamanho de pétala, largura da pétala, tamanho de sépala e largura de sépala. O PCA foi utilizado para reduzir a dimensionalidade do conjunto de dados para possibilitar a visualização das informações em um gráfico de duas dimensões, sendo os eixos denominados por *PC 1* (*principal component 1*) e PC 2 (*principal component 2*).

Os dados originais (definidos no espaço original) são mapeados para o espaço de baixa dimensão (definido pelas duas componentes principais mais relevantes), conforme os valores dos atributos e os coeficientes dos autovetores obtidos pelo PCA. Após esse mapeamento, pode-se observar alguns padrões de valores no espaço das componentes principais, por exemplo, quanto maior o valor de PC1, maior a largura e comprimento das pétalas. Nesse sentido, as flores do tipo setosa (pontos azuis) estão afastadas das flores do tipo

versicolor (pontos verdes) e *virginica* (pontos vermelhos), podendo-se concluir que as primeiras possuem características que as diferem dos outros dois tipos, tais como a largura e o comprimento da pétala, por terem valores menores de PC1.

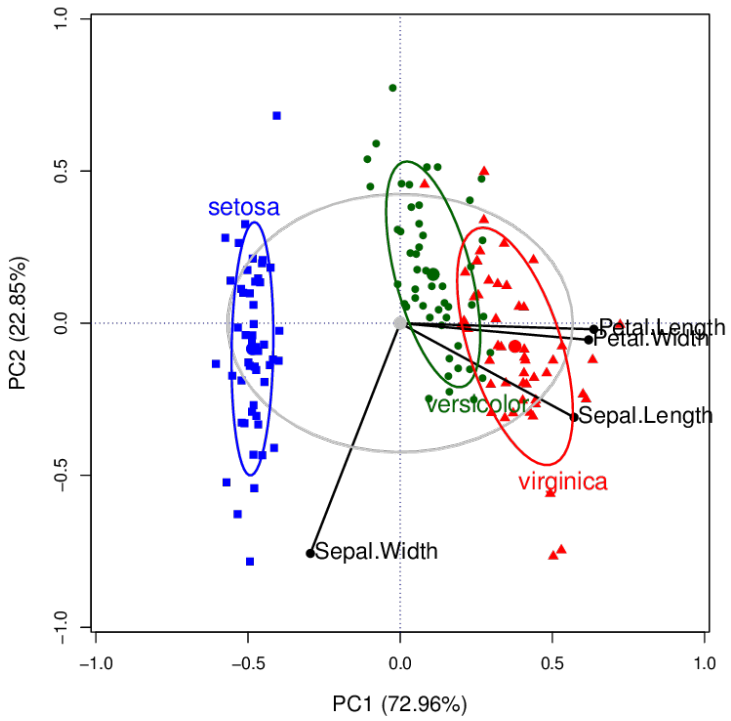


Figura 2.3: Exemplo de aplicação utilizando PCA.

t-Distributed Stochastic Neighbor Embedding

O t-SNE é um algoritmo de redução de dimensionalidade particularmente indicado para visualizações multidimensionais não-lineares. Diferente do PCA, esta métrica utiliza uma medida de similaridade, a distância Euclidiana, para "aprender" discrepâncias entre os pares de instâncias de dados. Desta forma, a estrutura e os padrões dos dados são preservados.

O algoritmo t-SNE é derivado do *Stochastic Neighbor Embedding* (SNE) que converte as distâncias Euclidianas de alta dimensão entre dados semelhantes em probabilidades condicionais. Diferente do SNE, o t-SNE usa a distribuição t-Student ao invés da Gaussiana para computar a similaridade [49].

A Figura 2.4 apresenta uma visualização do conjunto de dados *Íris*, gerada pelo algoritmo t-SNE. Da mesma forma que o PCA, uma visualização baseada no posicionamento de pontos no espaço bidimensional foi produzida empregando o t-SNE para reduzir a dimensionalidade dos dados. Observa-se que existem diferenças entre as visualizações geradas pelo PCA e pelo t-SNE, uma vez que há diferença nas suas formulações matemáticas

ao transformar os dados multidimensionais em um espaço de dimensões reduzidas. Tais diferenças podem ser vistas na geometria dos grupos de pontos, dados pelas espécies de flores, como também pela dispersão desses pontos no interior desses grupos.

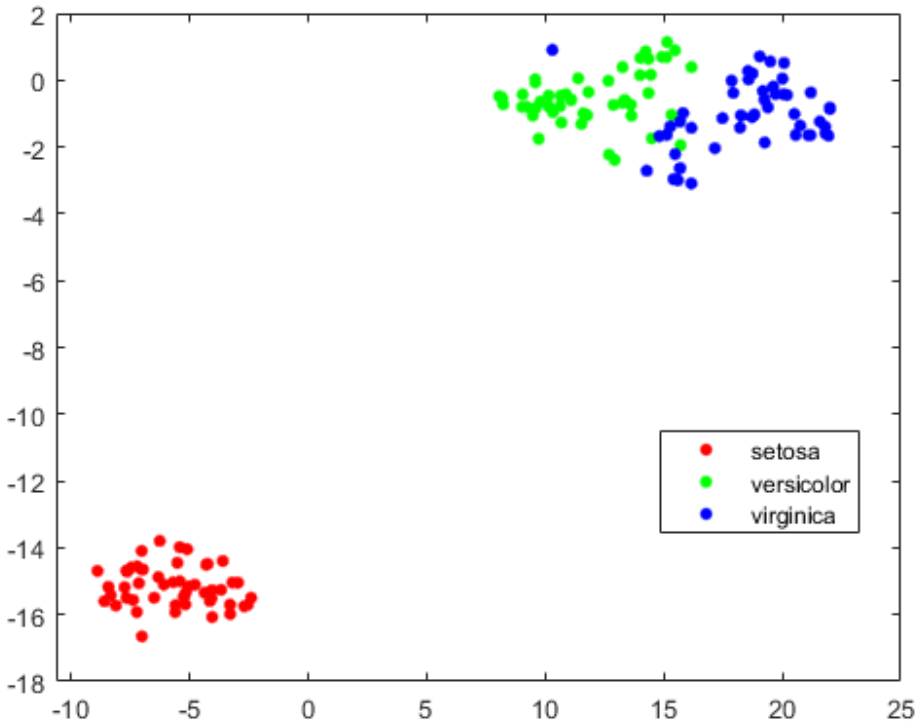


Figura 2.4: Visualização do conjunto de dados Íris utilizando o algoritmo t-SNE.

Capítulo 3

Revisão de Literatura

Este capítulo revisa os trabalhos literários que utilizaram temáticas referentes ao presente estudo. A Seção 3.1 trata de artigos sobre mulheres na computação, a Seção 3.2 dos projetos sobre mineração de dados educacionais aplicando técnicas de visualização de informação. Por último, a Seção 3.3 explana a relevância deste trabalho.

3.1 Mulheres na Computação

Na literatura é possível encontrar diversos trabalhos que estudam as razões de existirem poucas mulheres atuantes nas áreas de computação [50]. Outros trabalhos buscam encontrar as melhores formas de atrair mais pessoas do sexo feminino para cursos na referida área [51].

Em Lagesen [51] foram analisadas as quatro estratégias de inclusão para recrutar mulheres para o curso de Ciência da Computação. Tais estratégias foram: alcançar uma massa crítica, reforma educacional, redefinir o simbolismo de gênero da ciência da computação e mudar o conteúdo da disciplina. Os resultados demonstraram que aumentar o número de mulheres recrutadas é a principal estratégia para mudar a percepção simbólica do curso tornando-a menos "masculina" e mais neutra. Além disso, o aumento de mulheres no curso parece causar uma melhora no ambiente de aprendizado, já que problemas de minoria (visibilidade demais e atenção indesejada) se tornam menos presentes.

Já em Nunes [50] foi realizado um mapeamento de iniciativas brasileiras que estimulam a entrada ou a permanência de mulheres na computação. Utilizando um método próprio, semelhante ao método do Mapeamento Sistemático (MS), tendo como etapa - definição das questões de pesquisa e busca das iniciativas, definição dos critérios de inclusão e exclusão e classificação das iniciativas e análise destas. Foram encontradas as seguintes iniciativas brasileiras relevantes: *Women in Information Technology* (WIT); Fórum Meninas Digitais; *Android Smart Girls*; Meninas.comp; Projeto da UNIPLAC; Mulheres na Tecnologia;

Roda da Hacker; T.I.mosia; Emilias armação em bits; Feminino livre; *Girl Geek Dinners Brazil*; Inspirada na Computação; Meninas na engenharia; Projeto da UFMT-Mulheres na Computação; Projetomeninasmaismais; e Blog mulheres na Computação.

Com a análise do tipo de iniciativa foi possível verificar que *workshops* e palestras são os mais comuns devido ao caráter dinâmico e prático. Já os agenciamentos e *blogs*, os que possuem menos quantidade, devido ao fato de serem mantidos por uma pessoa que informa a respeito do que se encontra sobre o tema.

3.2 Mineração de Dados Educacionais

Existem ainda outros trabalhos que utilizam dados educacionais em todo o mundo, com o objetivo de aumentar a performance e o progresso dos alunos, entender a motivação, prever quais alunos irão abandonar o curso, além dos trabalhos que auxiliam o instrutor a entender os tópicos de discussão e a saber quais alunos estão participando. Os trabalhos fazem uso de algoritmos de visualização e/ou predição.

Em Asif [52] é analisada a performance de estudantes de graduação em cursos de Bacharelado em Tecnologia de Informação no Paquistão. O artigo teve como objetivo apresentar informações sobre o desempenho desses alunos, tanto para os professores envolvidos quanto para os diretores do programa de estudos, com vistas a ajudá-los a melhorar o programa, com o apoio da base de pesquisa de mineração de dados educacionais: predição, clusterização e destilação dos dados para julgamento humano.

A pesquisa foi separada em três questões: possibilidade de prever a performance dos estudantes com acurácia razoável no início da graduação; identificar matérias que servem como indicador de boa ou má performance do aluno; e identificar o progresso do aluno durante seus estudos, relacionado com o desempenho nas matérias. Para a primeira questão, foi possível prever a performance da graduação durante os quatro anos de programa, com base nas menções obtidas nos primeiros dois anos. O resultado foi de acurácia numérica, possibilitando, ao utilizar a técnica de árvore de decisão, observar quatro matérias que impactavam na segunda questão. Para solucionar a última questão, observou-se que os alunos com baixo desempenho no início do curso tendem a continuar com o mesmo desempenho. O estudo concluiu que é necessário um impacto maior em alunos que obtiveram o desempenho pior durante a graduação, por meio de aplicação de políticas, tão logo seja observado a situação no início do curso.

Em Domíngues [53] é apresentado o projeto SPEET (*Student Profile for Enhancing Engineering Tutoring*) que busca determinar e categorizar os diferentes perfis de engenharia na Europa para aprimorar ações de tutoria e auxiliar os alunos a obterem melhores resultados e concluírem a graduação. O trabalho empregou técnicas de visualização e

análise de dados utilizando os projetos SPEET e ERASMUS+ para promover suporte na tomada de decisão da tutoria ao analisar a performance dos alunos. Neste trabalho foi implementado um protótipo que utiliza histogramas coordenados e redução de dimensionalidade interativa. A visualização das coordenadas auxilia os educadores a entenderem a real influência da natureza (obrigatória ou eletiva) e da metodologia dos cursos (prática ou teórica), além de visualizar a mobilidade e a distribuição das notas. Já a redução de dimensionalidade foi empregada no reconhecimento de padrões nos dados.

Em Vieira et. al [39] é realizada uma busca por abordagens, propósitos, contextos e fonte de dados utilizados em análises visuais de aprendizado, além de procurar entender como são integradas, na literatura, teorias educacionais com princípios de visualização e análise visual de aprendizado.

O artigo concluiu que foram poucos os estudos realizados em salas de aula, devido ao fato de que a utilização de tecnologias para obter dados permite a coleta de um número maior. Desta forma, existe ainda a necessidade de se explorar a análise visual de aprendizado no contexto de sala de aula, por ser este um ambiente mais controlado. Os estudos também mostraram que os pesquisadores e os educadores analisaram a contribuição e a interação dos dados de forma isolada de outras informações relevantes na interpretação, tais como performance, informações demográficas, entre outros. Por mais que os alunos sejam os principais atores da produção dos dados e os instrutores a principal audiência para as ferramentas de visualização, observa-se, contudo, que existem diversas oportunidades para que os alunos aproveitem as ferramentas de análise de aprendizado visual e promovam o desenvolvimento de atividades cognitivas e metacognitivas.

Ainda no artigo, para realizar uma análise da integração dos temas na literatura, foram utilizadas duas técnicas de visualização - coordenadas paralelas e gráfico de dispersão. Desta forma, foi identificada uma falha na confluência entre teorias educacionais, visualizações sofisticadas e práticas de visualização de informação, na qual nenhum dos estudos analisados obteve avaliação máxima em todas as três dimensões exploradas.

Baker [54] apresenta a existência de diversos métodos dentro da mineração de dados educacionais que, normalmente, pertencem a cinco categorias: predição, clusterização, mineração de relacionamento, descoberta com um modelo, e destilação do dado para julgamento humano.

Predição tem como objetivo desenvolver o modelo que pode inferir um aspecto singular do dado a partir da combinação de outros aspectos. Requer rótulos que representam informações confiáveis sobre o valor da variável de saída dentro um conjunto de dados limitados. Clusterização objetiva encontrar pontos de dados que se agrupam naturalmente, dividindo o conjunto de dados completo em um conjunto de *clusters*. Mineração de relacionamento tem como meta descobrir o relacionamento entre as variáveis de um grande

conjunto de dados. Na descoberta com um modelo é desenvolvido um modelo de um fenômeno via previsão, agrupamento ou, em alguns casos, engenharia de conhecimento. Neste último caso o desenvolvimento usa o conhecimento humano e não o método automático.

Por fim, os dados são repassados para o julgamento humano com os objetivos de identificação e classificação. Na destilação para informação, os dados são dispostos de maneira a possibilitar o ser humano identificar facilmente os padrões implícitos, difíceis de expressar formalmente. Na destilação para classificação, subseções de um conjunto de dados são exibidas em formato visual ou de texto, e rotuladas por codificadores humanos.

3.3 Considerações Finais

A literatura é bem extensa no âmbito de estratégias e iniciativas para inclusão de um maior número de mulheres na área de tecnologia. Porém, com base na mineração de dados, percebe-se que existem diversas áreas a serem exploradas que podem auxiliar a identificar as razões da baixa adesão de mulheres, bem como o alto índice de abandono. Desta forma, as técnicas avançadas de visualização integrada com teorias educacionais e mineração de dados serão fundamentais para a construção deste trabalho.

Capítulo 4

Metodologia da Análise

Este capítulo apresenta todos os processos necessários para o levantamento das hipóteses do trabalho. Na Seção 4.1, são apresentados todos os passos para descoberta do conhecimento. Ele é dividido em Subseção 4.1.1, que apresenta os dados específicos que serão utilizados dentro da pesquisa; Subseção 4.1.2, que busca retirar os dados ruidosos presentes; Subseção 4.1.3, que trabalha com o armazenamento do dado para facilitar no uso de mineração de dados; Subseção 4.1.4, que utiliza as técnicas apresentadas nos dados trabalhados. Por fim, na Seção 4.2 são levantadas as hipóteses que este trabalho buscar responder.

4.1 Processo de Descoberta do Conhecimento

A tecnologia de informação disponibiliza meios para coleta, armazenamento e tratamento dos dados, sendo útil para sistemas educacionais que buscam avaliar as informações para melhor tomada de decisão. Para tanto, softwares e ferramentas, aliados a uma metodologia bem definida, fornecem suporte adequado ao processo decisório.

Destas ferramentas de tomada de decisão, destaca-se a Mineração de Dados, também conhecida como *Data Mining*, que é um processo minucioso de extração que busca padrões não evidentes em consultas às bases de dados. A mineração é uma das etapas do Processo de Descoberta de Conhecimento em Bases de Dados, ou KDD, *Knowledge Discovery in Databases*, que busca a extração não trivial de conhecimento previamente desconhecido, e potencialmente útil de um banco de dados [55].

O KDD é composto por cinco fases: Seleção, Limpeza, Transformação, Mineração de Dados e Avaliação. As etapas são executadas de forma interativa e iterativa, por envolver a cooperação da pessoa responsável pela análise de dados, e pelo fato de que esse processo não é aplicado de forma sequencial, mas envolve repetidas seleções de parâmetros [56]. A Figura 4.1 ilustra as fases do Processo de Descoberta do Conhecimento. Entretanto,

para este trabalho, a metodologia foi adaptada e o processo de mineração de dados foi substituído pela visualização dos dados.

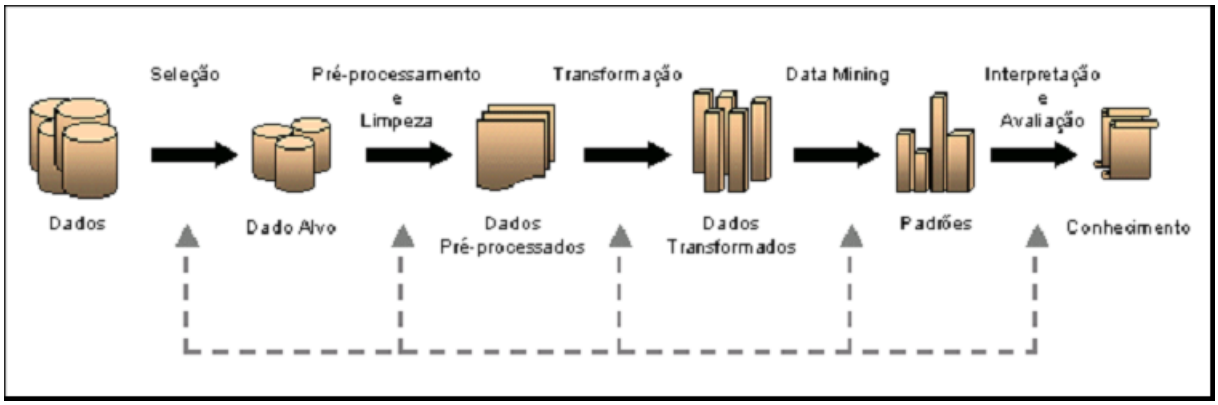


Figura 4.1: Metodologia Original do Processo de Descoberta do Conhecimento.

4.1.1 Seleção dos Dados

A fase de Seleção de Dados envolve a seleção de itens específicos para o Processo de Descoberta do Conhecimento. Nesta primeira fase foram selecionados apenas os dados do Sistemas Acadêmicos da Universidade de Brasília (SIGRA), com informações sobre os alunos dos cursos de tecnologia, sendo eles: Ciência da Computação, Licenciatura da Computação, Engenharia de Computação, Engenharia de Controle de Automação (Engenharia Mecatrônica), Engenharia de Redes de Comunicação e Engenharia de Software.

As informações obtidas foram divididas em três categorias: informações sobre os alunos, informações sobre as matérias, e informações de cada aluno sobre as matérias. Em relação aos alunos têm-se as seguintes variáveis:

- Sexo: informação se é feminino ou masculino;
- Data de nascimento: data no formato AAAA-MM-DD;
- UF de nascimento: sigla da Unidade da Federação onde o aluno nasceu;
- Cotista: sim, caso seja cotista, ou não, caso não seja;
- Tipo de escola: qual escola o aluno estudou antes de entrar na UnB, escola pública, escola particular ou não informada;
- Raça: relacionada à cor da pele, sendo as opções, preta, parda, branca, amarela, não cadastrada, não informada e não dispõe de informação;

- Curso: curso do aluno, sendo possível Engenharia Mecatrônica, Engenharia de Software, Engenharia de Redes de Comunicação, Engenharia de Computação, Computação e Ciência da Computação;
- Opção: opção escolhida pelo aluno ao ingressar na UnB, sendo, Engenharia Mecatrônica, Engenharia de Software, Engenharia de Redes de Comunicação, Engenharia de Computação, Computação e Ciência da Computação;
- Período de ingresso na UnB: período em que o aluno entrou na UnB, podendo ser desde o primeiro semestre de 1991 até o segundo semestre de 2016;
- Período de ingresso na opção: período em que o aluno ingressou na opção, podendo ser desde o primeiro semestre de 1991 até o segundo semestre de 2016;
- Forma de ingresso na UnB: como o aluno ingressou na universidade, tendo como opção: vestibular, transferência obrigatória, transferência facultativa, Sistema de Seleção Unificada (SISU), Programa de Avaliação seriada (PAS), portador de diploma de curso superior, matricula cortesia, ENEM, acordo cultural (PEC-G) ou convênio, sendo Int ou Andifes;
- Período de saída da opção: período em que o aluno deixou a opção podendo ser desde o primeiro semestre de 1992 até as férias de verão de 2017;
- Forma de saída da opção: forma como o aluno saiu da opção, sendo: novo vestibular, desligamento por abandono, formatura, desligamento por não cumprir condição, desligamento voluntário, mudança de curso, reprovação três vezes na mesma disciplina obrigatória, desligamento por decisão judicial, transferência, falecimento, desligamento por força de Convênio, aluno ativo, vestibular para outra Habilitação, desligamento por jubilamento, mudança de turno, desligamento por falta de documentação, desligamento por força de intercâmbio ou outros;
- Mínimo de créditos para formatura: número de créditos necessários para que o aluno se forme no curso;
- Créditos total cursados: número de créditos cursados pelo aluno até período da obtenção dos dados;
- Créditos integralizados no total: número de créditos que já foram integralizados;
- Créditos a integralizar: número de créditos que falta integralizar.

Sobre as matérias há as seguintes variáveis:

- Nome: nome da matéria no sistema Matrícula Web;

- Créditos: número de créditos concedidos por concluir a matéria;

As informações sobre os alunos a respeito das matérias são:

- Média no período: média da turma no período em que o aluno cursou a matéria;
- Nota: nota do aluno na matéria;
- Ano e semestre: período em que o aluno cursou a matéria;
- Créditos cursados no semestre: número de créditos cursados no semestre em que o aluno esteve matriculado na matéria;

A primeira categoria analisada se refere as informações de perfil dos alunos, entretanto, as informações sobre a relação aluno-matéria também são importantes e devem ser consideradas para uma análise futura.

4.1.2 Limpeza dos Dados

A fase de Limpeza dos Dados, também chamada de pré-processamento, busca corrigir as inconsistências nos dados. Desta forma, foram realizadas análises em linguagem R para verificar falta de informação, redundância, inconsistência e ruídos [57].

Assim, foram verificados se existiam valores infinitos ou valores inválidos, e investigada a presença de valores que não correspondiam ao esperado, tendo sido encontrados dois alunos que cursaram uma matéria de 60 créditos. A referida matéria é "estágio de internato 1 em medicina geral" e "estágio de internato 2 em medicina geral". Após uma análise mais aprofundada dos dois alunos, foram encontradas diversas matérias de medicina cursadas por eles na modalidade créditos concedidos. Disto, pode-se inferir que esses alunos vieram de outra escola de medicina e tiveram os créditos aproveitados para os cursos de Ciência da Computação e Licenciatura em Computação.

Na fase de Limpeza dos Dados também foram realizadas correções de divergências de código de caracteres; unificação como Engenharia Mecatrônica das opções Engenharia de Controle e Automação e Engenharia Mecatrônica, por serem o mesmo curso; além de transformar as letras das variáveis para minúscula. Por último, agregou-se na opção "Saiu" as seguintes condições: novo Vestibular, os diversos modos de desligamento, mudança de curso, transferência, vestibular para outra habilitação, mudança de turno e reprovação três vezes na mesma disciplina obrigatória, reduzindo o escopo de variáveis da forma de saída do curso.

4.1.3 Enriquecimento

Na Fase de Enriquecimento os dados são armazenados de forma estruturada para facilitar o uso das técnicas de mineração. Neste trabalho, os dados foram armazenados em um banco de dados relacional mySQL, por se tratar de uma estrutura simples.

As tabelas "Alunos", "Matérias" e "Matéria_aluno" foram criadas com as informações de perfil dos alunos, detalhes das matérias e relações do aluno com a matéria, respectivamente. A modelagem é ilustrada na Figura 4.2 e os atributos de cada tabela foram detalhadas na Seção 4.1.1.

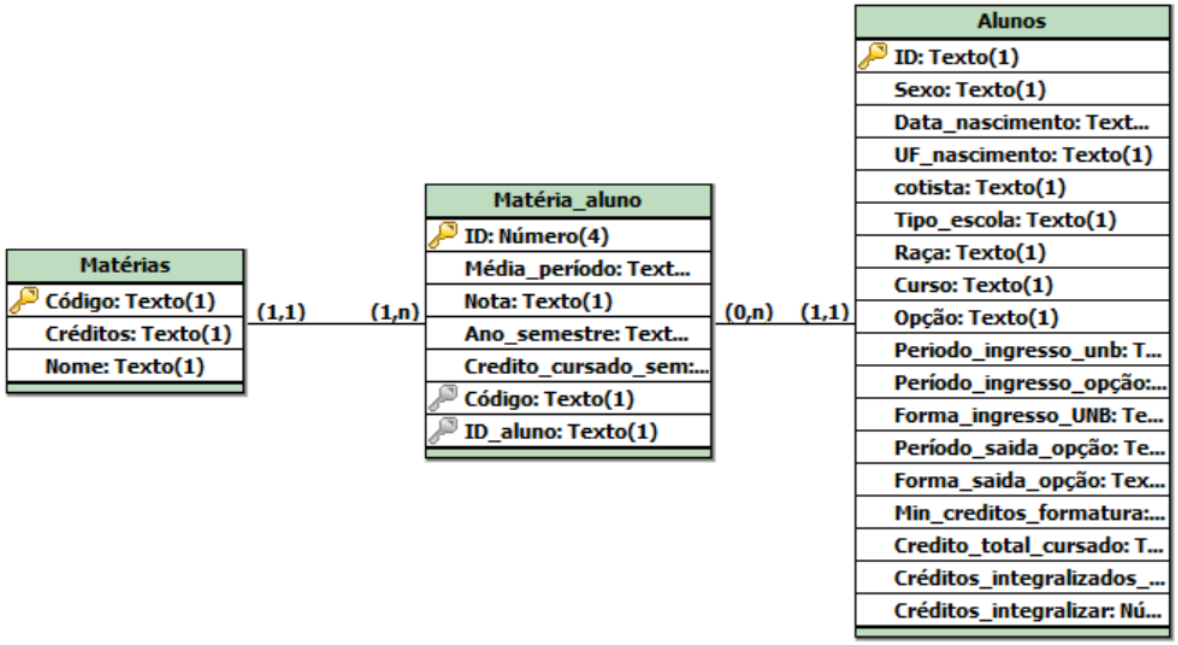


Figura 4.2: Modelo lógico do banco de dados criado para armazenar os dados dos alunos nos cursos de computação.

Após a Fase de Enriquecimento, foram selecionados os alunos do sexo feminino para realização de análises mais profundas, dando continuidade para a Fase de Visualização, uma vez que o foco desta pesquisa é estudar os perfis de alunas nos cursos de Computação. Ao final restou um total de 789 alunas, a Tabela 4.1 ilustra o número de alunas por curso

4.1.4 Visualização de Dados

A fase de Visualização de Dados, considerada a mais importante, é aquela, após o pré-processamento, na qual os dados contarão somente com as informações relevantes à pesquisa a ser feita. Nesta fase os algoritmos de visualização são utilizados para descobrir padrões relevantes nos dados.

Tabela 4.1: Número de indivíduos por curso ao final da limpeza dos dados.

Curso	**Número de alunas**
Computação	117
Ciência da Computação	205
Engenharia Mecatrônica	118
Engenharia de Redes de Comunicação	196
Engenharia de Software	90
Engenharia de Computação	63

Para encontrar as relações de similaridade implícitas nos dados, foram utilizadas técnicas de redução de dimensionalidade (PCA e t-SNE) e algoritmos de estudo de variáveis (HeatMap de variáveis e Coordenadas Paralelas).

O algoritmo HeatMap evidencia as relações entre contextos e fontes de dados. Esse algoritmo facilita a extração de valores numéricos exatos ou bits únicos de informação, mesmo dentro de grandes conjuntos de dados. Já a técnica Coordenadas Paralelas tem o propósito de facilitar o entendimento de tendências complexas e identificar pequenas discrepâncias de forma mais eficiente e mais compacta. Por último, a vantagem dos diagramas de dispersão, como o PCA e o t-SNE, reside na capacidade de representar diretamente relações bivariadas (por exemplo, dependência, associação, outliers) entre variáveis distintas, onde as lacunas podem ser facilmente identificadas.

4.2 Hipóteses

Das informações contidas na tabela "Alunos", deseja-se descobrir questões relacionadas ao perfil das alunas que estudam em cursos de tecnologia, entre elas:

1. As cotas são utilizadas pelas alunas de tecnologia?
2. As escolas podem influenciar na forma de ingresso ou no curso escolhido pelas alunas?
3. A região onde a aluna vive pode influenciar no curso escolhido?
4. As alunas que saem sem concluir o curso estão longe de se formar ou faltam poucos créditos?
5. É possível identificar padrões de perfil que diferem alunas: ativas, formadas e que evadiram do curso?

As perguntas foram elaboradas buscando entender o perfil das alunas e se há diferenças que influenciam na conclusão dos cursos. Para responder as questões 1, 2 e 3 foi necessário analisar a relação entre as variáveis cotas, tipo de escola, forma de ingresso e curso. Para a

questão 4 foi necessário investigar a relação entre o número de créditos e a forma de saída. Por último, o item 5 requer a utilização da redução de dimensionalidade para verificar graficamente se é possível diferenciar os grupos.

Capítulo 5

Resultados

Neste capítulo são apresentados os resultados obtidos a partir dos dados coletados e trabalhados, com foco na resolução das hipóteses levantadas no Capítulo 4. Para isso, foram utilizadas técnicas de visualização tradicionais, como *Heatmap* e Coordenadas paralelas, e técnicas de visualização, baseadas em redução de dimensionalidade, como *PCA* e *t-SNE*. Estas técnicas foram utilizadas para explorar os dados e encontrar padrões implícitos, uma vez que os dados apresentam muitas dimensões.

O PCA se caracteriza por ser uma técnica linear, enquanto que o t-SNE consiste em técnica mais complexa, capaz de reduzir a dimensionalidade em dados de natureza não-linear. Métodos lineares são computacionalmente mais eficientes, devido ao fato de que o mapeamento de cada instancia para visualização do espaço cartesiano, numa transformação linear, é obtido pela multiplicação de matriz e vetor. Entretanto, métodos lineares não conseguem resolver estruturas complexas em uma visão espacial, tornando-os propensos a introduzir distorções consideráveis ao lidar com variáveis não padronizadas [58]. Ambas as técnicas foram aplicadas e os resultados são demonstrados a seguir.

5.1 *Principal Component Analysis*

Componentes Principais são uma combinação linear normalizada das variáveis originais de um conjunto de dados. Na análise destes componentes, os problemas indagados podem ser abordados pelos seguintes objetivos [59]:

1. Identificação dos fatores principais;
2. Estudo das variáveis para seleção das mais pertinentes, cujo objetivo é reduzir o trabalho de obtenção de dados em estudos posteriores;
3. Análise dos dados no espaço reduzido, num sistema de coordenadas independentes (as duas ou três primeiras Componentes Principais).

5.1.1 Identificação dos Principais Fatores

Na técnica PCA os autovalores e os autovetores são informações importantes para interpretar padrões existentes nos dados e em seus atributos. A Figura 5.1 ilustra que o primeiro componente principal é o que concentra a maior parte da variância dos dados. O segundo, terceiro e quarto componentes, concentram proporções de variância menores, mas que juntas explicam grande parte da variabilidade dos dados. Na análise gráficas dos resultados do PCA, foram consideradas as 10 primeiras Componentes Principais, uma vez que a décima primeira concentra menos de 1% de variância nos dados.

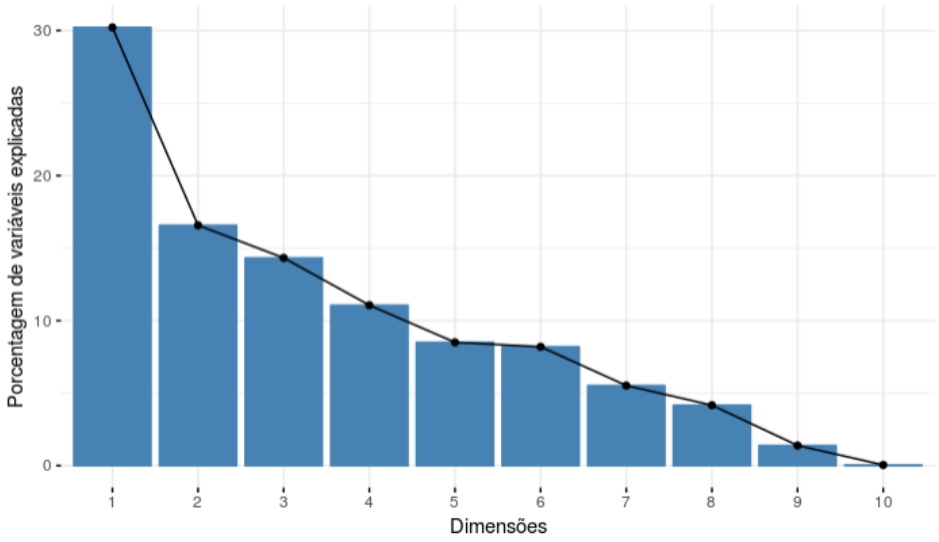


Figura 5.1: Porcentagem de variância explicada por cada componente principal.

Tabela 5.1: Sete primeiras componentes principais.

PC	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Des.Padrão	3.323	1.823	1.575	1.2168	0.935	0.901	0.607
Variância (%)	30.214	16.57	14.32	11.062	8.502	8.197	5.523
Var.Acumulada (%)	30.214	46.788	61.109	72.171	80.67	88.87	94.39

O campo "variância" representa a variabilidade dos dados e a "variância acumulada" representa a soma dessas variâncias, ambas em porcentagem. As Tabelas 5.1 e 5.2 descrevem que a variância acumulada nas 7 primeiras componentes principais explica 94% da variabilidade dos dados. Se optássemos por utilizar essas componentes, estaríamos reduzindo o número de 11 para 7 variáveis latentes, perdendo menos de 6% da informação da variabilidade dos dados.

Para entender a relação das variáveis com cada componente, foi utilizado o mapa bidimensional, também chamado de perceptual, explanado na próxima seção. Logo, sabendo que cada autovetor está associado com um autovalor, foi definido um espaço formado pelas

Tabela 5.2: Quatro últimas componentes principais.

PC	8	9	10	11
Des.Padrão	0.458	0.153	0.0052	2.5e-31
Variância	4.166	1.391	0.047	2.3e-30
Var.Acumulada	98.56	99.95	100	100

duas componentes principais (PCs) dado pelos autovetores relacionados aos autovalores, que concentram a maior variância dos dados.

5.1.2 Estudo das Variáveis

O mapa perceptual é representado por um diagrama que descreve a relação dos elementos de uma mesma categoria com diferentes características [60]. Vale ressaltar que a visualização do mapa perceptual pode aproximar alguns atributos levando em consideração a correlação entre eles, como também sua contribuição e padrão de variação em relação às componentes.

A Figura 5.2 representa o mapa perceptual, que ilustra a contribuição de cada variável nas dimensões 1 e 2 das Componentes Principais, sendo o eixo horizontal a dimensão 1 (Dim1), e o eixo vertical a dimensão 2 (Dim2). É possível observar quatro quadrantes no gráfico, onde o primeiro quadrante tem as dimensões 1 e 2 com valores positivos e o terceiro quadrante com valores negativos. O segundo quadrante apresenta as dimensões 1 e 2, respectivamente, positiva e negativa, já o quarto quadrante representa o inverso.

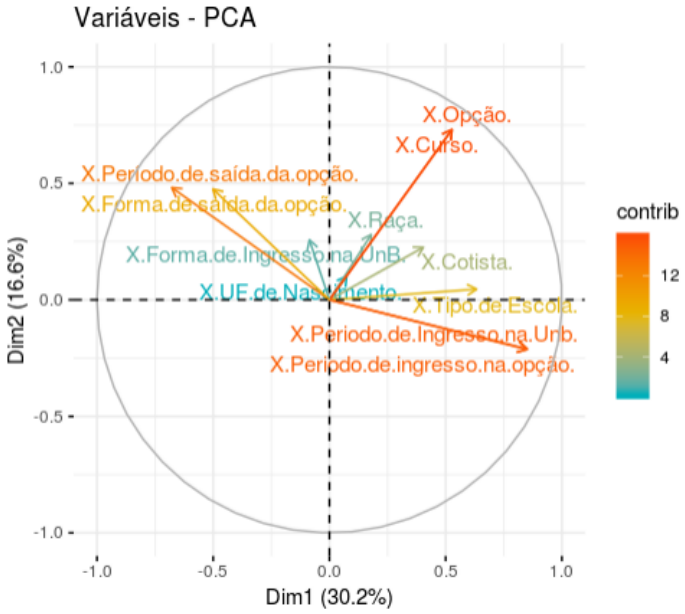


Figura 5.2: Gráfico de variáveis utilizando a técnica Mapa Perceptual.

No mapa, as variáveis que possuem a mesma direção têm correlação positiva, enquanto que as variáveis que possuem direções opostas têm correlação negativa. Ao considerar duas variáveis positivamente correlacionadas X e Y, se os valores de X aumentam, os valores de Y também aumentam. Por outro lado, se os valores de X diminuem, os valores de Y acompanham. Agora, se X e Y são duas variáveis negativamente correlacionadas, se os valores de X aumentam, então os valores de Y diminuem e vice-versa. Desta forma, dentre as variáveis analisadas, as que possuem correlação positiva são:

- período de ingresso na universidade e o período de ingresso na opção;
- o curso e a opção;
- forma de saída da opção e período de saída da opção.

Vale ressaltar que o período de ingresso na universidade é diferente do período de ingresso na opção, visto que o aluno tem a possibilidade de realizar uma mudança de curso, onde é mantida a matrícula. De modo semelhante, o período de ingresso na opção e na universidade possuem correlação negativa com o período de saída da opção.

Utilizando os valores de contribuições das variáveis que foram representadas na Figura 5.2, foi gerada a Tabela 5.3. As Tabelas 5.4 e 5.5 representam os maiores valores das contribuições nas componentes principais 1 (Dim1) e 2 (Dim2), respectivamente.

Tabela 5.3: Contribuição das variáveis para as dimensões 1 e 2.

VARIÁVEIS	**DIMENSÃO 1**	**DIMENSÃO 2**
UF.de.Nascimento	0.07420169	0.1008586
Cotista	0.40429806	0.2255627
Tipo.de.Escola	0.63305415	0.0457119
Raça	0.17972664	0.2798621
Curso	0.52720613	0.7305449
Opção	0.52720613	0.7305449
Periodo.de.Ingresso.na.Unb	0.85208999	-0.2114406
Periodo.de.ingresso.na.opção	0.84897207	-0.2129773
Forma.de.Ingresso.na.UnB	-0.08618908	0.2570947
Período.de.saída.da.opção	-0.67833241	0.4815968
Forma.de.saída.da.opção.	-0.50133795	0.4755097

Com relação a Dimensão 1 observa-se que, quanto maior o valor das variáveis período de ingresso (tanto na UnB quanto na opção), tipo de escola, curso e opção, maior será o valor da primeira Componente Principal. Altos valores na Dimensão 1 são relacionados a baixos valores no período de saída e na forma de saída, por serem correlacionados negativamente.

Tabela 5.4: Principais variáveis que contribuem para a dimensão 1.

VARIÁVEIS	DIMENSÃO 1	DIMENSÃO 2
Periodo.de.Ingresso.na.Unb	0.85208999	-0.2114406
Periodo.de.ingresso.na.opção	0.84897207	-0.2129773
Tipo.de.Escola	0.63305415	0.0457119
Curso	0.52720613	0.7305449
Opção	0.52720613	0.7305449
Período.de.saída.da.opção	-0.67833241	0.4815968
Forma.de.saída.da.opção.	-0.50133795	0.4755097

Na dimensão 2 é possível perceber que as variáveis com os maiores valores de contribuição são: curso, opção, período de saída e forma de saída que variam de forma conjunta. A segunda Componente Principal é fortemente influenciada por valores correlacionados positivamente.

Tabela 5.5: Principais variáveis que contribuem para a dimensão 2.

VARIÁVEIS	DIMENSÃO 1	DIMENSÃO 2
Curso	0.52720613	0.7305449
Opção	0.52720613	0.7305449
Período.de.saída.da.opção	-0.67833241	0.4815968
Forma.de.saída.da.opção.	-0.50133795	0.4755097

Buscando uma melhor visualização da correlação entre as variáveis, utilizou-se a técnica *Heatmap*, explicada na Seção 2.3.1, para ilustrar o nível de correlação entre cada variável, seja positiva ou negativa. Na Figura 5.3 é representado o mapa de calor da matriz de correlação, onde as cores de cada correlação indicam a força e o sinal desta, estando em vermelho as negativas e em azul as positivas.

Para localizar as variáveis com maior grau de relacionamento, as mais correlacionadas foram agrupadas em N grupos, gerando assim o gráfico da Figura 5.4. O uso do Heatmap facilita interpretação da análise das correlações, pois possibilita a associação de cores aos diferentes graus de correlação entre variáveis e a realização de agrupamentos dos atributos com base nas correlações.

Na Universidade de Brasília existe um sistema de cotas que reserva uma quantidade de vagas de 5% para negros, de ambos os tipos de escolaridade; e 50% para estudantes de escolas públicas. Dentro dos últimos 50%, 25% são destinados aos estudantes com renda familiar bruta per capita igual ou inferior a um salário e meio, já os outros 25%, aos alunos com renda superior. Dentro da divisão das cotas de escolas públicas, uma parte das vagas é reservada aos alunos que se declaram pardos, pretos ou indígenas. Esta porcentagem é definida com base na soma total da população que integra esses grupos em cada unidade da Federação, conforme o último censo do IBGE.

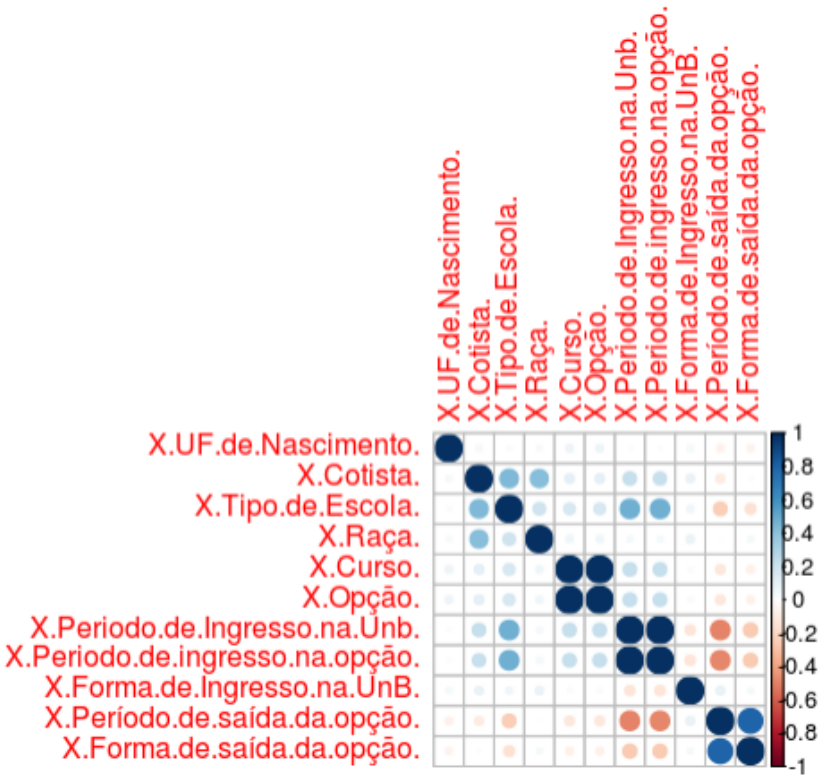


Figura 5.3: Mapa de calor representando a correlação das variáveis.

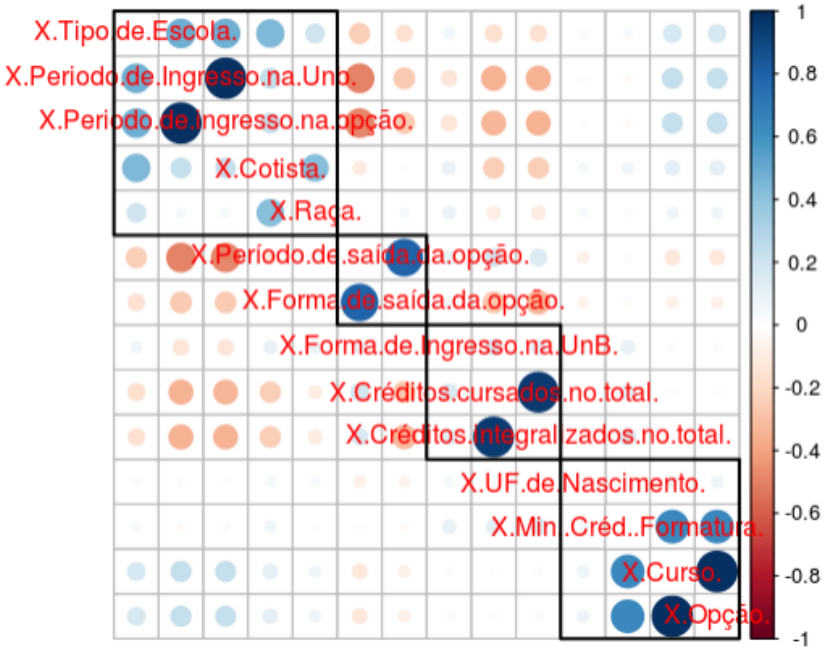


Figura 5.4: Mapa de calor agrupado pelas variáveis mais correlacionadas.

Analisando o grupo étnico no mapa de calor, é possível perceber que a variável cotista tem correlação positiva com o período de ingresso na UnB, uma vez que o sistema de cotas para negros na UnB entrou em vigor há 15 anos, o que corrobora com a correlação obtida entre esses atributos. Outras variáveis que variam juntas por possuírem correlação positiva estão relacionadas aos atributos cotas, raça, tipo da escola e o período de ingresso na opção.

Observa-se, no mapa de calor, que o curso possui correlação positiva com o tipo de escola (escola pública ou particular), indicando assim que a escolha do curso é influenciada pela escola que a aluna cursou o ensino médio. Outra correlação positiva observada se refere ao mínimo de créditos para a formatura com o curso, já que para se formar o aluno precisa concluir o valor mínimo dos créditos, que variam conforme o curso.

Como uma das hipóteses deste trabalho é investigar as razões de saída das alunas dos cursos de Computação, é interessante analisar o atributo relacionado com a forma de saída dessas alunas. De acordo com a Figura 5.4, pode-se notar que a forma de saída possui correlação positiva com o período de saída e correlação negativa tanto com o tipo de escola quanto com o período de ingresso na UnB e na opção. Nesse sentido, os dados referentes a forma de saída do curso foram classificados como: "Formadas", sendo as alunas que terminaram o curso, "Ativas", as alunas que ainda estão cursando, e "Saiu", as alunas que saíram do curso por algum motivo que não seja formatura.

A Figura 5.5 apresenta o resultado da aplicação da técnica Coordenadas Paralelas, empregada com o intuito de analisar as relações entre o período de entrada na opção e a situação final das alunas, conjugada com o atributo de forma de saída da opção. Percebe-se um padrão nas poli-linhas do gráfico associado às alunas que saíram do curso em dois períodos: entre o primeiro semestre de 1992 até o segundo semestre de 1995, e entre o primeiro semestre de 2001 até o segundo semestre de 2016. Em relação às alunas que formaram, observa-se um padrão de poli-linhas no período do primeiro semestre de 1996 até o primeiro semestre de 2000 e do segundo semestre de 2007 até as férias de verão de 2017. Por fim, as ativas são aquelas que não possuem período de saída, uma vez que as poli-linhas conectam ao valor zero no período de saída da opção.

Na Figura 5.6 é possível perceber a relação das variáveis analisadas anteriormente (forma de saída e período de saída) com o período de ingresso, tanto na UnB, quanto na opção. Desta forma, percebe-se que as alunas ativas entraram na opção, no período de 2007 a 2016, já aquelas que formaram, entraram na opção entre primeiro semestre de 1991 e o segundo semestre de 2013. Curiosamente, pode-se observar alguns casos de alunas formadas que entraram no segundo semestre de 2015. Tendo em vista que, os cursos de computação possuem, no mínimo, 8 semestres, supõe-se que tais alunas vieram de outros cursos e já tinham créditos cursados, possibilitando a formatura em 5 semestres.

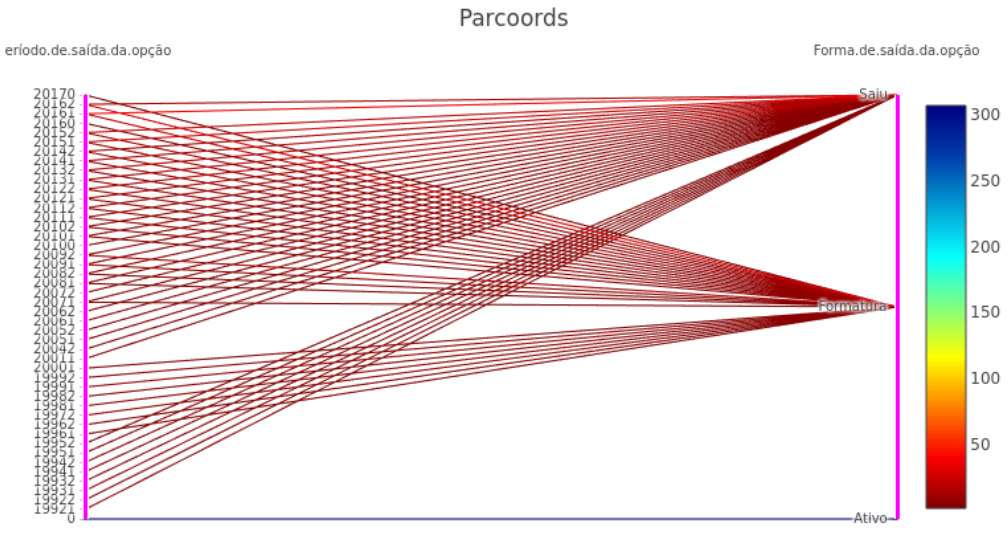


Figura 5.5: Relação entre forma de saída e período de saída.

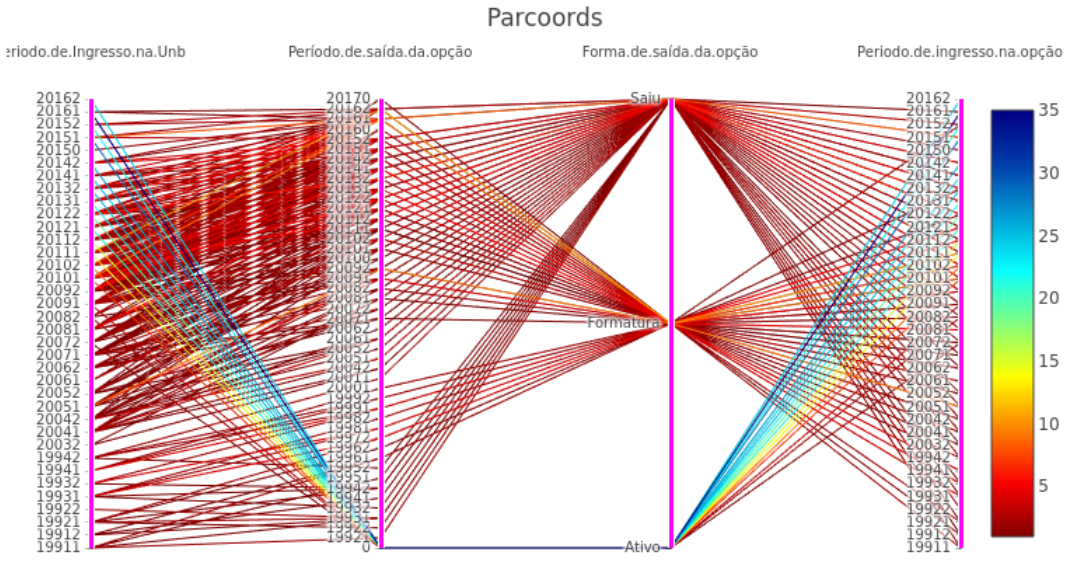


Figura 5.6: Relação entre forma de saída, período de saída e períodos de entrada.

Buscando diferenciar as alunas que estão em diferentes grupos de forma de saída, a próxima seção apresenta uma visualização de dados em espaço de dimensionalidade reduzida de acordo com os padrões da matriz de correlação.

5.1.3 Visualização de Pontos do PCA

A Figura 5.7 apresenta uma visualização baseada no posicionamento de pontos do espaço reduzido, sendo definido pelas duas componentes principais obtidas pelo PCA, Dim1 e

Dim2. Nessa visualização, cada ponto está associado com uma instância no conjunto de dados (registro de uma aluna), que está definido em espaço de alta dimensionalidade. As instâncias foram coloridas de acordo com o valor do atributo categórico "Forma de saída", pois a estratégia é encontrar perfis de meninas, considerando as situações "Ativas", "Formadas" ou "Desligadas".

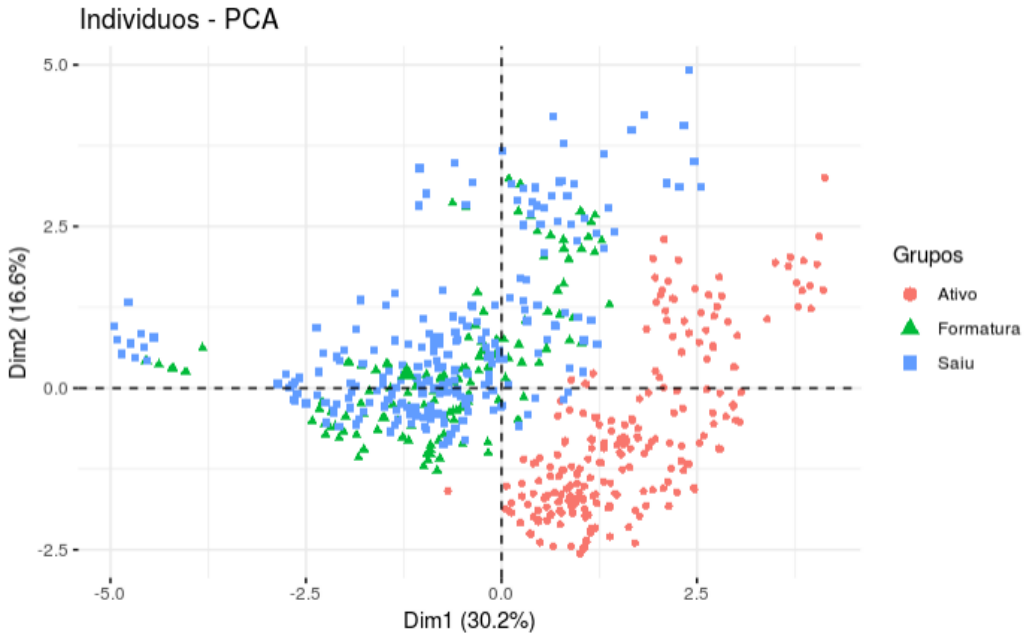


Figura 5.7: Gráfico de agrupamento de indivíduos por forma de saída utilizando a técnica PCA.

A análise da visualização da Figura 5.7 permite observar alguns padrões de acordo com a proximidade das instâncias em relação às cores associadas com as três formas de saída das alunas. Nesse sentido, pode-se notar que as alunas que "formaram" (pontos representados pela cor verde) e aqueles que não concluíram o curso (cor azul) têm características semelhantes, pois estão associados aos pontos posicionados na mesma região do gráfico, chegando a se sobrepor. Enquanto que as alunas ativas (cor vermelha) possuem características distintas, devido a delimitação de uma área no gráfico onde está concentrada a maior parte destes pontos.

Visando obter informações adicionais a respeito de cada categoria de alunas, as médias das coordenadas de cada grupo foram calculadas e apresentadas na Tabela 5.6. Percebe-se as coordenadas dos grupos "Formado" e "Saiu" bem próximas, estando os indivíduos do grupo "Saiu" ligeiramente mais acima que os indivíduos do grupo "Formado".

Tabela 5.6: Média dos centroides.

GRUPOS	DIMENSÃO 1	DIMENSÃO 2
Ativo	1.5544733	-0.8141180
Formado	-1.2102926	0.2199083
Saiu	-0.8298706	0.7358166

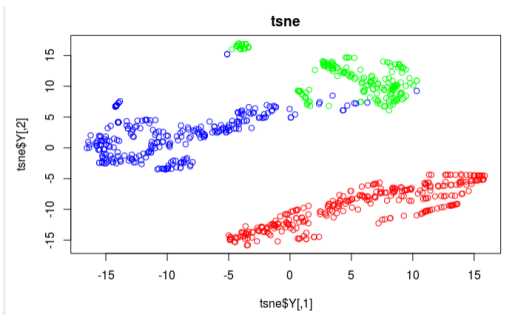
5.2 t-Distributed Stochastic Neighbor Embedding

Diferente do PCA, o t-SNE é um algoritmo de redução de dimensionalidade não-linear, baseado na distribuição de probabilidade com caminhos aleatórios em grafos de vizinhança para descobrir estruturas dentro dos dados. Devido às propriedades do algoritmo do t-SNE, quando realizadas várias execuções com os mesmos parâmetros, é possível a produção de resultados diferentes. Desta forma, para definir os melhores parâmetros para os dados utilizados, foram realizadas cinco repetições do algoritmo.

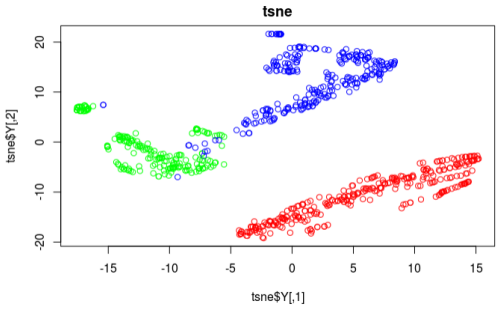
A Figura 5.8 apresenta as cinco repetições realizadas utilizando o algoritmo t-SNE com os parâmetros que geraram o melhor resultado. É possível perceber que realmente há alterações nas visualizações obtidas, entretanto, mantém-se um padrão na geometria dos grupos de pontos.

Na visualização obtida utilizando o PCA, foi possível identificar dois grupos: alunas ativas e alunas que saíram do curso (formadas ou desligadas). Ao passo que, com o t-SNE foi possível diferenciar os três grupos definidos, ativas, formadas e as que saíram do curso sem concluí-lo. Desta forma, é possível dizer que o algoritmo t-SNE gerou visualizações mais apropriadas para analisar as características de alunas do que o PCA. Isto posto, pode-se concluir que os dados possuem natureza não-linear e distribuição estatística peculiar, levando o algoritmo t-SNE a identificar meninas em diversos perfis.

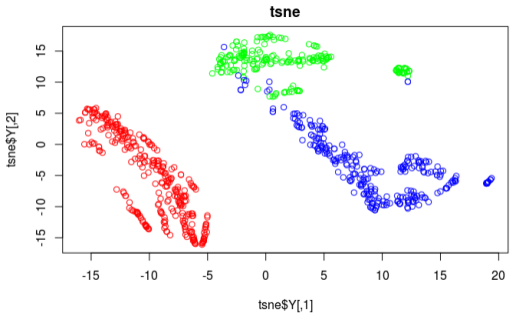
Com o intuito de descobrir características e padrões nos grupos gerados pelo t-SNE, é necessário recursos de interação da visualização com o usuário. Esta parte será discutida na Seção 6.2 de Trabalhos Futuros.



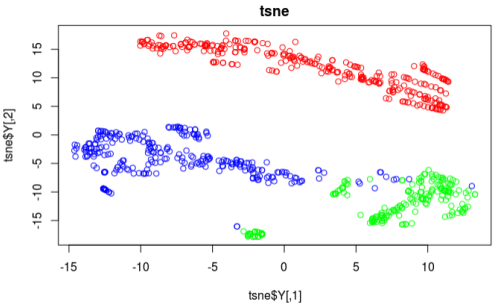
(a) Primeira execução



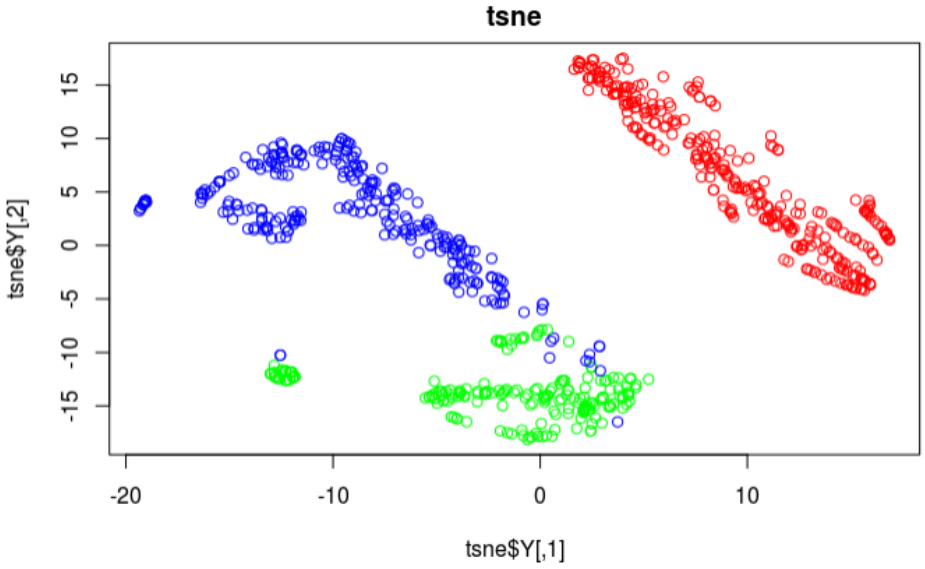
(b) Segunda execução



(c) Terceira execução



(d) Quarta execução



(e) Quinta execução

Figura 5.8: Cinco execuções do algoritmo t-SNE com os mesmos parâmetros.

Capítulo 6

Conclusão

6.1 Considerações Finais

Este trabalho consistiu na realização de uma análise visual dados das meninas na área de tecnologia dos cursos de graduação da Universidade de Brasília. Para tal, foi necessário coletar os dados considerados nesta pesquisa, realizar a limpeza desses dados, aplicar técnicas capazes de extrair informações imprescindíveis para desenvolver um maior entendimento das principais dificuldades enfrentadas pelas meninas nos cursos, transformando o resultado em conhecimento relevante.

Dentro da visualização da informação, a visualização exploratória pode ser entendida como um processo de geração de hipóteses, permitindo ao usuário obter informações para obtenção de conhecimento implícito e relevante. Para auxiliar neste trabalho foram escolhidas três técnicas de visualização tradicional - coordenadas paralelas e *heatmap* - e duas técnicas de visualização de dados multivariados - *PCA* e *t-SNE*.

Além disso, foi realizada uma análise dos dados de alunas de diferentes cursos de computação da Universidade de Brasília. A abordagem empregada considerou os atributos e as alunas agrupadas pela forma de saída da opção, obtendo-se informações relevantes do perfil dessas alunas. A análise visual dos dados das meninas levou em consideração as técnicas de visualização baseadas em *PCA* e *t-SNE*, em que foram geradas representações gráficas de diferentes tipos, como o gráfico de dispersão, mapa de calor e histograma.

As visualizações obtidas evidenciaram as correlações entre as variáveis, e agruparam as meninas de acordo com a forma de saída da universidade (ativas, formadas ou desligadas). Assim, foi possível determinar aspectos pertinentes sobre o perfil das alunas, tais como, a relevância das cotas para aquelas que se encaixam no perfil e o papel da escola na escolha do curso. Logo, comparando os resultados da redução de dimensionalidade, pode-se concluir que os dados são de natureza não-linear devido a melhor segregação pelo algoritmo t-SNE.

Pretende-se, com o presente trabalho, proporcionar aos departamentos dos cursos de computação da Universidade de Brasília a oportunidade da tomada de decisões no auxílio ao estabelecimento de políticas que proporcionam ambientes de ensino mais prazerosos, acessíveis e inclusivos para as meninas. Buscando, assim, incrementar o número de formandas, bem como o número de mulheres no mercado de trabalho.

6.2 Trabalhos Futuros

Para dar continuidade a este trabalho, são sugeridos os seguintes estudos:

- O estudo de outras técnicas de visualização de dados não-lineares juntamente com um processo de visualização exploratória para extrair informações implícitas e relevantes, concentrando esforços nas variáveis correlacionadas com a forma de saída;
- Utilização dos dados de cada aluna relacionada com as matérias para uma análise mais precisa, avaliando não apenas o perfil, mas também o comportamento na universidade, tais como as notas de cada matéria e a média do período em que esta foi cursada;
- Utilização da visualização como apoio às tarefas de mineração dos dados, podendo prever e analisar o desempenho das alunas de computação, distinguindo entre concluintes e não-concluintes.

Referências

[1] Santos, CM: *Por que as mulheres "desapareceram" dos cursos de computação?"*. Jornal da USP, Disponível em http://jornal. usp. br/universidade/por-que-as-mulheresdesapareceram-dos-cursos-de-computacao, publicado em, 7, 2018. ix, 3

[2] KANNO, Mário: *Marcos na história da visualização de dados*. Infografe. Disponível em:< http://euclid. psych. yorku. ca/SCS/Gallery/milestone/historia_infografia. pdf> Acessado em, 2, 2008. 1

[3] Freitas, Carla Maria Dal Sasso, Olinda Mioka Chubachi, Paulo Roberto Gomes Luzzardi e Ricardo Andrade Cava: *Introdução à visualização de informações*. Revista de informática teórica e aplicada. Porto Alegre. Vol. 8, n. 2 (out. 2001), p. 143-158, 2001. 1

[4] Estivalet, Luiz Fernando: *O uso de ícones na visualização de informações*. 2000. 1

[5] Baker, Ryan, Seiji Isotani e Adriana Carvalho: *Mineração de dados educacionais: Oportunidades para o brasil*. Brazilian Journal of Computers in Education, 19(02):03, 2011. 2

[6] *Educação superior em computação, estatísticas 2016*. Sociedade Brasileira de Computação-SBC. Disponível em:< http://www. sbc. org. br/documentos-da-sbc/summary/133-estatisticas/1074-educacaosuperior-em-Computação-estatisticas-2016>. Acesso em, 7, 2016. 2

[7] Holanda, Maristela; Araújo, Aleteia; Walter Maria Emília T.; Suertegaray Azucena; de Oliveira; Carlos Alberto Jesus: *Meninas.comp: Um relato da experiência de integração entre alunas e docentes do ensino médio e da universidade de brasília*. Em *CLEI*, 2018. 2

[8] Henn, Steve: *When women stopped coding*. NPR Planet Money, 21, 2014. 2

[9] Ferreira, Aurélio Buarque de Holanda. Em *Novo dicionário Aurélio da língua portuguesa*. 2004. 3

[10] Menabrea, Luigi Federico e Ada Lovelace: *Sketch of the analytical engine invented by charles babbage*, 1842. 3

[11] Lovelace, Ada King: *Ada, the Enchantress of Numbers: The Letters of Lord Byron's Daughter and Her Description of the First Computer*. Strawberry Press, 1992. 3

[12] Gurer, Denise W: *Pioneering women in computer science*. Communications of the ACM, 38(1):45–55, 1995. 3

[13] Fritz, W Barkley: *The women of eniac*. IEEE Annals of the History of Computing, 18(3):13–28, 1996. 3

[14] Ensmenger, Nathan: *Making programming masculine*. Gender codes: Why women are leaving computing, páginas 115–141, 2010. 3

[15] Maciel, Cristiano, Sílvia Amélia Bim e Karen da Silva Figueiredo: *Digital girls program-disseminating computer science to girls in brazil*. Em *2018 IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering (GE)*, páginas 29–32. IEEE, 2018. 4

[16] CNPq. Edital 18/2013 MCTI/CNPq/SPM-PR/Petrobras – Meninas e Jovens Fazendo Ciência Exatas, Engenharias e Computação. 4

[17] Unibanco, Instituto. Edital Gestão Escolar para Equiadae: Elas nas Exztas, 2015. Disponível em: http://www.fundosocialelas.org/elasnasexatas/. Acessado em 09 de Abril de 2017. 4

[18] Silberschatz, Abraham, Henry F Korth, Shashank Sudarshan *et al.*: *Database system concepts*, volume 4. McGraw-Hill New York, 1997. 6

[19] Silberschatz, Abraham, Henry F Korth, Shashank Sudarshan *et al.*: *Database system concepts*. 6:1–29, 2010. 6

[20] Ackoff, Russell L: *From data to wisdom*. 6

[21] Bellinger, Gene, Durval Castro e Anthony Mills: *Data, information, knowledge, and wisdom*. 2004. 7

[22] Rowley, Jennifer: *The wisdom hierarchy: representations of the dikw hierarchy*. Journal of information science, 33(2):163–180, 2007. 7

[23] Tan, Pang Ning *et al.*: *Introduction to data mining*. Pearson Education Índia, 2007. 7, 8

[24] Jain, Ramesh C, SN Jayaram Murthy, Luong Tran e Shankar Chatterjee: *Similarity measures for image databases*. Em *Storage and Retrieval for Image and Video Databases III*, volume 2420, páginas 58–66. International Society for Optics and Photonics, 1995. 7

[25] Spiegel, Murray R, John J Schiller, R Alu Srinivasan e Mike LeVan: *Probability and statistics*, volume 2. Mcgraw-hill New York, 2009. 8

[26] Rice, John A.: *Mathematical statistics and data analysis*. Wadsworth & Brooks/Cole, 2007. 9

[27] Morettin, Pedro Alberto e WILTON OLIVEIRA BUSSAB: *Estatística básica*. Editora Saraiva, 2017. 9

[28] Martins, Maria Eugénia Graça: *Coeficiente de correlação amostral*. Revista de Ciência Elementar, 2(2), 2014. 9

[29] Sedgwick, Philip: *Pearson's correlation coefficient*. Bmj, 345:e4483, 2012. 10

[30] Card, Mackinlay: *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999. 10

[31] Dias, Mateus Pereira: *A visualização da informação e a sua contribuição para a ciência da informação*. 10

[32] Tufte, Edward e P Graves-Morris: *The visual display of quantitative information.; 1983*, 2014. 10

[33] Dias, Maria Madalena, Juliana Keiko Yamaguchi, Emerson Rabelo e Clélia Franco: *Visualization techniques: Which is the most appropriate in the process of knowledge discovery in data base?* Em *Advances in Data Mining Knowledge Discovery and Applications*. InTech, 2012. 10

[34] Keim, Daniel A: *Information visualization and visual data mining*. IEEE Transactions on Visualization & Computer Graphics, (1):1–8, 2002. 10

[35] Ward, Matthew O, Georges Grinstein e Daniel Keim: *Interactive data visualization: foundations, techniques, and applications*. AK Peters/CRC Press, 2015. 11

[36] Deng, Wankun, Yongbo Wang, Zexian Liu, Han Cheng e Yu Xue: *Hemi: a toolkit for illustrating heatmaps*. PloS one, 9(11):e111988, 2014. 11

[37] YoshimiTanaka, Oswaldo, Marcos Drumond Júnior, Elier Broche Cristo, Sandra Maria Spedo e Nicanor Rodrigues da Silva Pinto: *Uso da análise de clusters como ferramenta de apoio à gestão no sus*. Saúde e Sociedade, 24:34–45, 2015. 11

[38] Wilkinson, Leland e Michael Friendly: *The history of the cluster heat map*. The American Statistician, 63(2):179–184, 2009. 11

[39] Vieira, Camilo, Paul Parsons e Vetria Byrd: *Visual learning analytics of educational data: A systematic literature review and research agenda*. Computers & Education, 122:119–135, 2018. 11, 18

[40] Keogh, Eamonn e Abdullah Mueen: *Curse of dimensionality*. Em *Encyclopedia of Machine Learning*, páginas 257–258. Springer, 2011. 13

[41] Campos, Teófilo Emıdio de: *Técnicas de seleção de características com aplicações em reconhecimento de faces*. Tese de Doutoramento, Master's thesis, Universidade de São Paulo, 2001. 13

[42] Souza, André Marcelo de, Ronei Jesus Poppi *et al.*: *Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes principais: um tutorial, parte i*. Química nova, 2012. 13

[43] Jolliffe, Ian T: *Principal component analysis and factor analysis*. Em *Principal component analysis*, páginas 115–128. Springer, 1986. 13

[44] Pearson, Karl: *Liii. on lines and planes of closest fit to systems of points in space*. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11):559–572, 1901. 13

[45] Fisher, Ronald A e Winifred A Mackenzie: *Studies in crop variation. ii. the manurial response of different potato varieties*. The Journal of Agricultural Science, 13(3):311–320, 1923. 13

[46] Hotelling, Harold: *Analysis of a complex of statistical variables into principal components.* Journal of educational psychology, 24(6):417, 1933. 13

[47] Wold, Svante, Kim Esbensen e Paul Geladi: *Principal component analysis*. Chemometrics and intelligent laboratory systems, 2(1-3):37–52, 1987. 13

[48] Duda, Richard O e Peter E Hart: *Pattern classification and scene analysis*. A Wiley-Interscience Publication, New York, 1973. 13

[49] Maaten, Laurens van der e Geoffrey Hinton: *Visualizing data using t-sne*. Journal of machine learning research, 9(Nov):2579–2605, 2008. 14

[50] Nunes, MASN, Carolina S Louzada, Edilayne M Salgueiro, Beatriz T Andrade, PS Lima e RMCT Figueiredo: *Mapeamento de iniciativas brasileiras que fomentam a entrada de mulheres na computação*. Em *Anais do XXXVI Congresso da Sociedade Brasileira de Computação-X Women in Information Technology (WIT 2016)*, páginas 2697–2701, 2016. 16

[51] Lagesen, Vivian Anette: *The strength of numbers: Strategies to include women into computer science*. Social Studies of Science, 37(1):67–92, 2007. 16

[52] Asif, Raheela, Agathe Merceron, Syed Abbas Ali e Najmi Ghani Haider: *Analyzing undergraduate students' performance using educational data mining*. Computers & Education, 113:177–194, 2017. 17

[53] Dominguez, Manuel, Ramon Vilanova, Miguel Prada, José Vicario, Marian Barbu, Maria João Pereira, Michal Podpora, Umberto Spagnolini, Paulo Alves e Anna Paganoni: *Speet: visual data analysis of engineering students performance from academic data*. LASI 2018-Learning Analytics Summer Institutes-Universidad de Leon, 2018. 17

[54] Baker, RSJD *et al.*: *Data mining for education*. International encyclopedia of education, 7(3):112–118, 2010. 18

[55] Costa, Claudio Napolis, Jonatas Vieira Coutinho, Lúcia Helena de Magalhães e Márcio Aarestrup Arbex: *Descoberta de conhecimento em bases de dados*. Revista Eletrônica: Faculdade Santos Dumont, 2. 20

[56] CORRÊA, ÂNGELA MC JORGE e HH Sferra: *Conceitos e aplicações de data mining*. Revista de ciência & tecnologia, 11:19–34, 2003. 20

[57] Navega, Sergio: *Princípios essenciais do data mining.* Anais do Infoimagem, 2002. 23

[58] Nonato, Luis Gustavo e Michael Aupetit: *Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment.* IEEE transactions on visualization and computer graphics, 2018. 27

[59] Duarte, João Batista: *Introdução à análise de componentes principais.* 27

[60] Santos, Glauber e Viviane Silva: *Mapa perceptual como ferramenta para a análise da imagem de destinos turísticos.* Revista de Turismo Contemporâneo, 3(2), dez 2015. `https://periodicos.ufrn.br/turismocontemporaneo/article/view/6856`. 29